63 /4/ 2025

JOURNAL OF THEORETICAL AND APPLIED MECHANICS

WARSAW 2025, QUARTERLY, VOLUME 63, ISSN 1429-2955 eISSN 2543-6309

POLISH SOCIETY OF THEORETICAL AND APPLIED MECHANICS ptmts

POLISH SOCIETY OF THEORETICAL AND APPLIED MECHANICS

JOURNAL OF THEORETICAL AND APPLIED MECHANICS

Vol. 63 • No. 4

Quarterly

WARSAW 2025

JOURNAL OF THEORETICAL AND APPLIED MECHANICS
(until 1997 Mechanika Teoretyczna i Stosowana, ISSN 0079-3701)

Beginning with Vol. 45, No. 1, 2007, *Journal of Theoretical and Applied Mechanics* (JTAM) has been selected for coverage in Thomson Reuters products and custom information services. Now it is indexed and abstracted in the following:

- **Science Citation Index Expanded** (also known as SciSearch®)
- **Journal Citation Reports/Science Edition**

Editorial Office

Al. Armii Ludowej 16, room 650; 00-637 Warsaw, Poland
e-mail: jtam@ptmts.org.pl
www.jtam.pl

Ministry of Science and Higher Education
Republic of Poland

This issue has been published with financial support from the Polish Ministry of Science and Higher Education under the Excellent Science II programme "Support for scientific conferences".

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 715–720, Warsaw 2025
https://doi.org/10.15632/jtam-pl/202666

PROBABILISTIC ESTIMATION OF THE DYNAMIC GAIT PARAMETERS†

Tomasz WALCZAK[1], Michał GUMINIAK[2*], Marcin KAMIŃSKI[3]

[1] *Institute of Applied Mechanics, Poznan University of Technology, Poznan, Poland*
[2] *Institute of Structural Analysis, Poznan University of Technology, Poznan, Poland*
[3] *Faculty of Civil Engineering, Architecture & Environmental Engineering, Lodz University of Technology, Lodz, Poland*

*corresponding author, *michal.guminiak@put.poznan.pl*

This study presents an estimation of the dynamic parameters of gait with a random approach. The data necessary for random analysis was obtained through laboratory tests. The study was conducted on a group of healthy people aged 20–25, without diagnosed musculoskeletal diseases. It consisted in walking along a several-metre-long path at free speed and recording the ground reaction forces (GRF) for both limbs using dynamometric platforms. On this basis, the basic dynamic parameters of gait, such as maxima and local minima of the stance phase were determined, and then they were subjected to stochastic analysis.

Keywords: dynamic gait parameters; semi-analytical probabilistic approach; stochastic perturbation technique; Monte Carlo simulation.

1. Introduction

Gait, as the basic form of locomotion, is the most frequently studied human activity. Walking properly is one of the distinguishing features of good health and condition of the examined person. Conversely, disturbed gait with abnormal parameters indicates potential problems with the musculoskeletal system, past injuries, joint dysfunctions, etc. Therefore, it is important to choose the right parameters that quantify the gait characteristics of the examined person. Among the parameters most frequently analyzed by researchers, a group of spatiotemporal and dynamic parameters can be distinguished. The parameters from the first group reflect the geometry and kinematics of the gait (spatial and kinematic parameters of the body and its segments), while the dynamic parameters reflect the forces and moments acting on the segments during walking. The most frequently analyzed ones include the characteristic course of ground reaction forces (GRF) measured during the limb support phase (Derlatka & Parfieniuk, 2023; Fryzowicz *et al.*, 2018; Michałowska *et al.*, 2018; Richards *et al.*, 2023). New indices are also defined to quantify walking behavior based on ground reaction force, e.g., (Park & Kim, 2022). Measured with dynamometer platforms or tensometric mats, they indicate whether the feet and joints of the lower limbs are correctly loaded, and the asymmetry occurring in the loads between the limbs. The most important parameters obtained during this type of research certainly include the parameters characterizing the course of the vertical component of the GRF, most often expressed in the percentage of the body weight of the examined person. Although it is generally known what the correct shape of the curve representing these reactions is, it depends on many factors such as body weight, gait speed, age, overall health of the person, potential dysfunctions of the musculoskeletal system, etc. Therefore, it is very difficult to determine in the studied group of

†The content of this article was presented during the 31st Conference Vibrations in Physical and Technical Systems – VIBSYS, Poznań, Poland, October 16–18, 2024.

people what a deviation from the norm is, despite the fact that the studied group is homogeneous, e.g., a group of athletes, people after a knee joint injury, or healthy people. The present paper proposes a probabilistic approach to gait analysis based on gait studies of a homogeneous group of healthy, young adults without diagnosed musculoskeletal diseases, where the relationship between key parameters describing the characteristics of the vertical component of GRF, depending on the mass of the examined person, was analyzed. Joint kinematics measurement plays the main role in describing the loads and kinematics of gait (Żuk & Trzeciak, 2017). A similar study on injuries of anterior cruciate ligament (ACL) in terms of stochastic approach was deeply examined (Lin *et al.*, 2012), where the Monte Carlo simulation (MCS) technique found application.

2. Formulation of the problem using a probabilistic approach

A finite number of deterministic solutions is necessary to carry out the probabilistic calculations. A Gaussian probability distribution of the observed design parameter is assumed. Due to the continuous distribution of the Gaussian probability density function $p_v(x)$, where x is the domain of occurrence of a given phenomenon, it is necessary to perform the continuous response function based on a finite number of deterministic results. Polynomial approximations using the least squares method (LSM) have been adopted to obtain response curves basing on a finite number of deterministic results. This allows us to determine the so-called system response fitting curves in the form of polynomials, using the LSM:

$$G = \sum_{j=0}^{n} C_j \cdot v^j. \tag{2.1}$$

This way, it is possible to express the probabilistic solution to each problem within the range determined by the coefficient of variation of the random design parameter. The semi-analytical method (SAM), the stochastic perturbation technique (SPT) and the MCS will be independently used to carry out the random analysis. Polynomials of the third order were adopted as fitting functions. The SAM is based on symbolic calculation procedures in the Maple program. All procedures of the SPT of the tenth order were carried out using the Maple program. Having this, probabilistic moments are calculated, i.e., the expectation (E), standard deviation (σ), coefficient of variation (α), skewness (β) and kurtosis (κ) (Kamiński, 2013; Kamiński *et al.* 2024):

$$\begin{aligned}
E(G) &= \int_{-\infty}^{+\infty} \sum_{j=0}^{n} C_{ij} v^j p_v(x)\,\mathrm{d}x, \\
\sigma(G) &= \left\{ \int_{-\alpha}^{\alpha} \left(\sum_{j=0}^{n} C_{ij} v^j - E[G] \right)^2 p_v(x)\,\mathrm{d}x \right\}^{1/2}, \\
\alpha(G) &= \left| \frac{\sigma(G)}{E(G)} \right|, \qquad \beta(G) = \frac{\mu_3(G)}{\sigma^3(G)}, \qquad \kappa(G) = \frac{\mu_4(G)}{\sigma^4(G)}.
\end{aligned} \tag{2.2}$$

MCS with the number of trials equal to 10^5 was carried out using the Maple program too, with the fact that probabilistic moments are calculated from statistical formulas. The primary objective of this study is to investigate the most important gait parameters in the examined group, employing a random approach based on weight.

3. Laboratory experiment

The study involved measuring the reaction forces of the ground as the test subject walked along a designated path at varying speeds in a natural gait. For this purpose, two dynamometric platforms, AMTI BP400600 with frequency of sampling 400 Hz, were used. An example of the vertical component of GRF with two characteristic maximum values: F_h – maximum force gained during heel strike, F_f – maximum force gained during terminal stance (forefoot press) for one limb during gait is shown in Fig. 1.

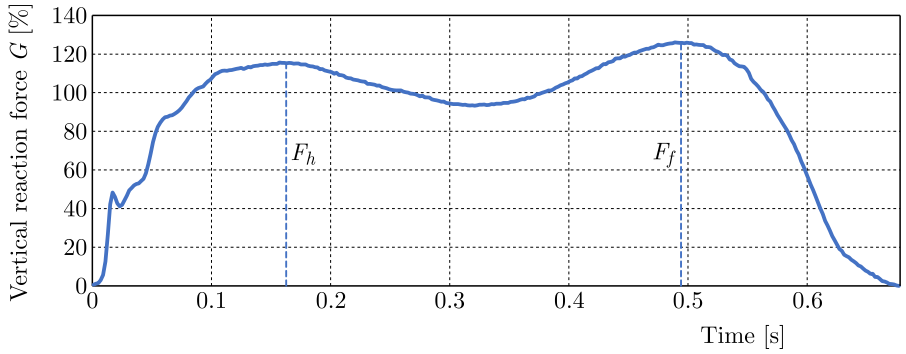

Fig. 1. Vertical component of GRF in % of body weight during gait.

The two maximum values presented in Fig. 1, characteristic of each vertical component of the GRF, were subjected to probabilistic analysis.

4. Probabilistic estimation of measured laboratory data

Loadings of the body under consideration were all found via the polynomial basis:

$$G = \sum_{j=0}^{n} C_j \cdot X^j. \tag{4.1}$$

4.1. Example 1 – heel strike maximum (F_h)

The response function is expressed as the third order polynomial:

$$G = -0.0659583377410756 - 1.20410359255101 \cdot X + 0.15008617465672 \cdot X^2 - 0.00167560644573114 \cdot X^3. \tag{4.2}$$

Figure 2 presents the results of the simulation of the expected values, kurtosis and skewness depending on the coefficient of variation, based on SPT, MSC and SAM methods.

Assuming that, according to the known rules, the desired distribution of a random variable is one in which the coefficient of variance is less than 5 %, we will notice that for the analyzed variable, this means an expected value within 110G for the studied group. In order for the distribution to be treated as normal, the kurtosis should have a value of 3 and the skewness should be as close to 0 as possible. It can be observed that in order for the conditions of normality of the distribution to be met, the deviation from the mean should not be greater than 2 %–3 %.

4.2. Example 2 – forefoot maximum press (F_f)

The response function is expressed as the third order polynomial:

$$G = 1.16382439132543 + 21.2823173261835 \cdot X - 0.664821019062622 \cdot X^2 + 0.00568578395391626 \cdot X^3. \tag{4.3}$$

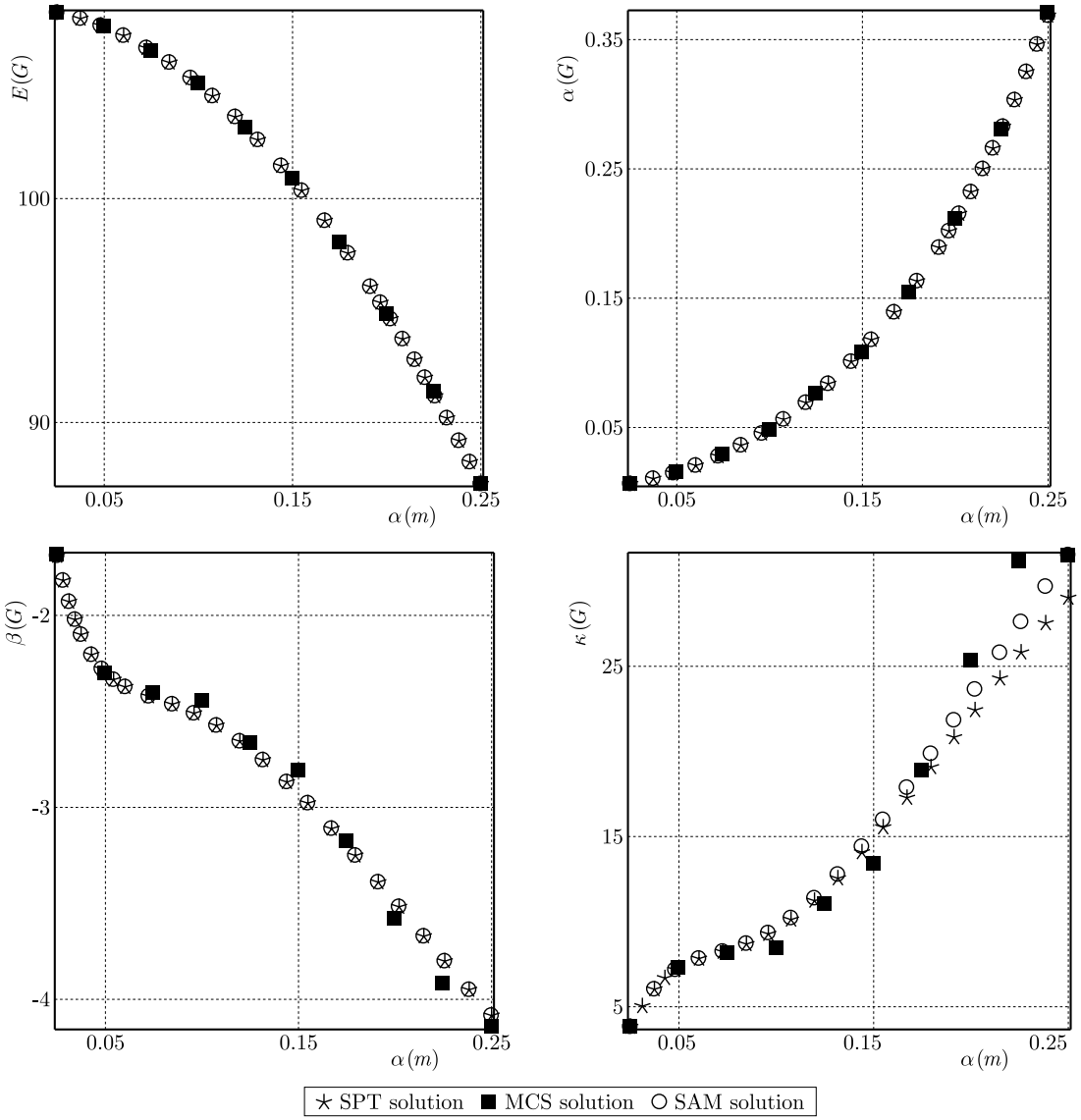

Fig. 2. Results of probabilistic calculations for the randomly distributed mass location.

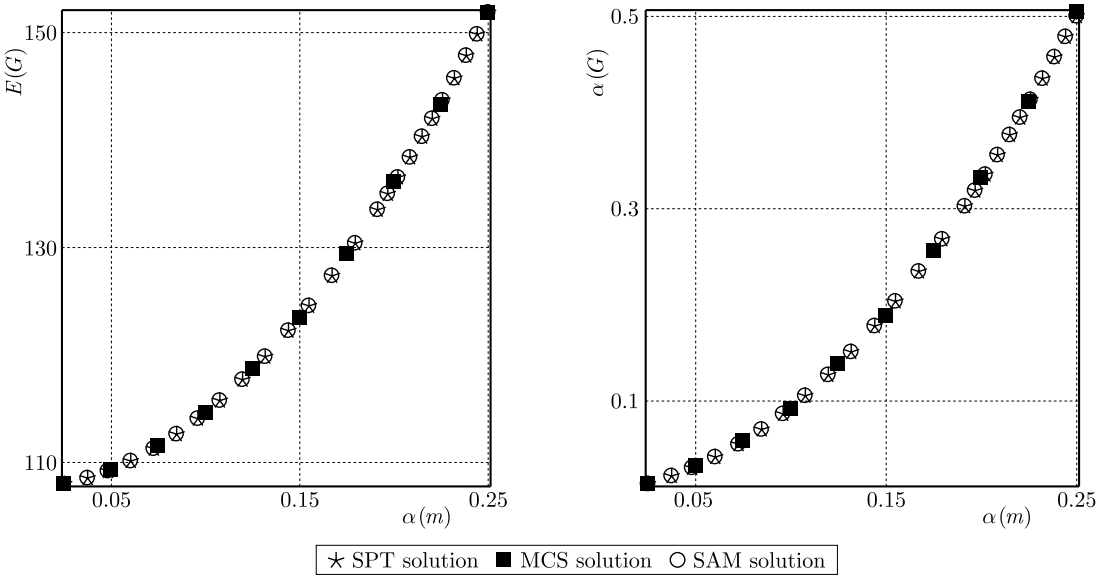

Fig. 3. Results of probabilistic calculations for the randomly distributed mass location.

Figure 3 presents, as an example, only the expected value depending on the coefficient of variation.

As in the case of the analysis of the first maximum, it can be noted that in order for the spread of the value of the analyzed variable not to be too large from the average, the expected value should oscillate within 110G. The other analyzed parameters have similar characteristics as in Fig. 2. It can be concluded that in order for the analyzed variables to have a normal distribution in the studied group, it would be necessary to remove those cases that generate "tails" in the distribution.

5. Conclusions

The calculation results presented in the previous chapter allow the following conclusions:

- laboratory experiments and computational studies presented in this work clearly demonstrate that the common application of three probabilistic approaches – SAM, SPT, and MCS techniques – enables the accurate determination of the probabilistic coefficients of external loading in the presence of input Gaussian uncertainty. In most cases, very good agreement has been noticed between these three methods. The simplest random approach is the semi-analytical one, and it allows us to derive random moments in analytical terms. It seems to be relatively easy for future implementations in much more complex biomechanical issues;
- the Monte Carlo method is characterized by a relatively long computation time depending on the number of trials. The real number of trials that can provide correct results is 100 000;
- the presented analysis can provide valuable statistical information on the parameters determined in the analyzed group of people, and will allow the assessment of whether statistical inference is justified;
- in biomedical research, a common situation is when the size of the research group is too small. The presented approach can be useful in a process of supplementing the data in such a situation;
- the method will allow observing and eliminating from the group units that statistically differ from the rest;
- the work is a contribution to further research and presents preliminary analyses of the random approach.

Acknowledgments

The presented research results were funded with the grant 0612/SBAD/3576 allocated by the Ministry of Science and Higher Education in Poland.

References

1. Derlatka, M., & Parfieniuk, M. (2023). Real-world measurements of ground reaction forces of normal gait of young adults wearing various footwear. *Scientific Data*, *10*, Article 60. https://doi.org/10.1038/s41597-023-01964-z
2. Fryzowicz, A., Murawa, M., Kabaciński, J., Rzepnicka, A., & Dworak, L.B. (2018). Reference values of spatiotemporal parameters, joint angles, ground reaction forces, and plantar pressure distribution during normal gait in young women. *Acta of Bioengineering and Biomechanics*, *20*(1), 49–57.
3. Kamiński, M. (2013). *The stochastic perturbation method for computational mechanics*. Wiley.
4. Kamiński, M., Przychodzki, M., Łasecka-Plura, M., Guminiak, M., & Lenartowicz, A. (2024). Random eigenvibrations of beams with viscoelastic layers. *Journal of Theoretical and Applied Mechanics*, *62*(4), 763–767. https://doi.org/10.15632/jtam-pl/194242

5. Lin, C.-F., Liu, H., Gros, M.T., Weinhold, P., Garret, W.E., & Yu, B. (2012). Biomechanical risk factors of non-contact ACL injuries: A stochastic biomechanical modeling study. *Journal of Sport and Health Science*, *1*(1), 36–42. https://doi.org/10.1016/j.jshs.2012.01.001
6. Michałowska, M., Walczak, T., Grabski, J.K., & Cieślak, M. (2018). People identification based on dynamic determinants of human gait. *Vibrations in Physical Systems*, *29*, Article 2018012.
7. Park, J.S., & Kim, C.H. (2022). Ground-reaction-force-based gait analysis and its application to gait disorder assessment: New indices for quantifying walking behavior. *Sensors*, *22*(19), Article 7558. https://doi.org/10.3390/s22197558
8. Richards, J., Levine, D., & Whittle, M.W. (2023). *Whittle's gait analysis* (6th ed.). Elsevier.
9. Żuk, M., & Trzeciak, M. (2017). Anatomical protocol for gait analysis: Joint kinematics measurement and its repeatability. *Journal of Theoretical and Applied Mechanics*, *55*(1), 369–376. https://doi.org/10.15632/jtam-pl.55.1.369

Manuscript received December 8, 2024; accepted for publication March 6, 2025; published online September 10, 2025.

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 721–736, Warsaw 2025
https://doi.org/10.15632/jtam-pl/203130

SENSITIVITY ANALYSIS OF MULTIPLE EIGENVALUES AND ASSOCIATED EIGENVECTORS OF QUADRATIC EIGENPROBLEM[†]

Henryk CIUREJ

Faculty of Civil Engineering and Resource Management, AGH University of Cracow, Cracow, Poland
hciurej@agh.edu.pl

The article is focused on the sensitivity of eigenvalues and eigenvectors in a quadratic eigenvalue problem with real matrices defining the problem under consideration and under the strong assumption that these matrices form a non-defective operator. The particular interest is the case of multiple eigenvalues and associated eigenvectors. Generally in such a case derivatives in the Fréchet sense do not exist, but only in the Gâteaux sense. The formulas of the directional differential in the closed matrix form were derived. A numerical example is shown.

Keywords: quadratic eigenvalue problem; multiple eigenvalues; sensitivity analysis.

1. Introduction

Conscious shaping of the dynamic characteristics of a structure is an essential duty of the designer, especially for structures where the dynamic phenomena require non-standard calculations and not just the use of dynamic coefficients as static load multipliers. Knowledge of the properties of the eigenproblem and its importance in the dynamics of the whole system is fundamental in such cases. Determining the sensitivity of the eigenvalues and eigenvectors can be helpful in structural optimization or identification. On the basis of the presented formulas, an original programme was developed that enables the calculation of the directional derivatives of eigenvalues and eigenvectors.

Sensitivity analysis related to the eigenproblem has been intensively developed for at least 80 years. On both the mathematical and numerical side, it is still a research problem – especially in terms of the numerical efficiency of the algorithms developed (Łasecka-Plura, 2023; 2024; Phuor & Yoon, 2023; Martinez-Agirre & Elejabarrieta, 2011; Wang & Dai, 2015). Indeed, the mathematical foundations and basic physical interpretations of the problems of sensitivity analysis of multiple eigenvalues and eigenvectors have been well known since the 1990s. In general, the majority of papers deals with a symmetric problem, i.e., when the matrices of a system of equations of motion in the configuration space are symmetric.

The linear eigenproblem (LEP) $(\mathbf{A} - \mathbf{I})\,\mathbf{I} = \mathbf{0}$, is extensively discussed in abundant mathematical literature, as well as in the literature on dynamical systems (Horn & Johnson, 2013; Garcia & Horn, 2017). The quadratic eigenproblem (QEP) is less frequently discussed due to the computational practice of reducing a QEP to a LEP via an isospectral transformation (Xu & Wu, 2008). It should be noted that QEPs and LEPs are nevertheless different because of the different spaces of the eigenvectors and their properties (Tisseur & Meerbergen, 2001; Lancaster & Zaballa, 2009; Lancaster, 2013).

[†]The content of this article was presented during the 31st Conference Vibrations in Physical and Technical Systems – VIBSYS, Poznań, Poland, October 16–18, 2024.

The method for deriving the derivatives presented in this paper combines and extends various concepts presented in (Seyranian *et al.*, 1994; Krog & Olhoff, 1995; Lee *et al.*, 1999b) – the combining element is the way in which the eigenvalues are numbered and ordered, and the extension pertains to the non-symmetry of matrices in the system of the equations of motion in configuration space.

It will be assumed in the rest of the text that $\mathbf{h} = \left[h_1, ..., h_{N_p}\right]$ denotes a parameter vector in the design parameter space, $\mathbf{h} \in \mathcal{R}^{N_p}$, N_p – dimension of the space. In that space, the directional versor $\mathbf{e} \in \mathcal{R}^{N_p}$, ($\|\mathbf{e}\| = 1$) is also defined. All quantities appearing in the QEP, therefore, remain dependent on the vector $\mathbf{h}$.

For convenience, let us introduce, according to (Andrew *et al.*, 1993; Lancaster, 2013; Lancaster & Zaballa, 2009), the operator $\mathbf{L}$ defined as follows:

$$\mathbf{L}\left(\lambda(\mathbf{h}), \mathbf{h}\right) = \lambda^2(\mathbf{h})\mathbf{M}(\mathbf{h}) + \lambda(\mathbf{h})\mathbf{C}(\mathbf{h}) + \mathbf{K}(\mathbf{h}), \tag{1.1}$$

where the dependence on the parameter vector is explicitly indicated. Then the QEP can be written in abbreviated form:

$$\mathbf{L}\left(\lambda(\mathbf{h}), \mathbf{h}\right)\boldsymbol{\Psi}(\mathbf{h}) = \mathbf{0}, \qquad \boldsymbol{\Phi}^H(\mathbf{h})\mathbf{L}\left(\lambda(\mathbf{h}), \mathbf{h}\right) = \mathbf{0}. \tag{1.2}$$

It follows that both eigenvalues and eigenvectors are mappings of the vector $\mathbf{h}$, i.e., $\lambda = \lambda(\mathbf{h})$, $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\mathbf{h})$, and $\boldsymbol{\Phi} = \boldsymbol{\Phi}(\mathbf{h})$ – in the general case, these mappings for multiple eigenvalues are no longer differentiable in the Fréchet sense, but only in the Gâteaux one.

It is assumed that the matrices occurring in the eigenproblem are dependent on the vector $\mathbf{h}$, i.e.:

$$\mathbf{M} = \mathbf{M}(\mathbf{h}), \qquad \mathbf{C} = \mathbf{C}(\mathbf{h}), \qquad \mathbf{K} = \mathbf{K}(\mathbf{h}), \tag{1.3}$$

and it is assumed also that these matrices are differentiable in the Fréchet sense, so it implies the existence of partial derivatives and the Taylor series expansion in the form:

$$\begin{aligned}
\mathbf{K}(\mathbf{h} + \epsilon\mathbf{e}) &= \mathbf{K}(\mathbf{h}) + \epsilon\sum_{p=1}^{N_p}\frac{\partial\mathbf{K}(\mathbf{h})}{\partial h_p}e_p, \\
\mathbf{C}(\mathbf{h} + \epsilon\mathbf{e}) &= \mathbf{C}(\mathbf{h}) + \epsilon\sum_{p=1}^{N_p}\frac{\partial\mathbf{C}(\mathbf{h})}{\partial h_p}e_p, \\
\mathbf{M}(\mathbf{h} + \epsilon\mathbf{e}) &= \mathbf{M}(\mathbf{h}) + \epsilon\sum_{p=1}^{N_p}\frac{\partial\mathbf{M}(\mathbf{h})}{\partial h_p}e_p,
\end{aligned} \tag{1.4}$$

where $\epsilon \in \mathcal{R}$.

The article is focused on the sensitivity of eigenvalues and eigenvectors in a QEP with real matrices defining the problem under consideration and under the strong assumption that these matrices form the non-defective operator $\mathbf{L}$ (i.e., the operator $\mathbf{L}$ is diagonalizable). The QEP is defined in classical form by the equation:

$$\left(\lambda^2\mathbf{M} + \lambda\mathbf{C} + \mathbf{K}\right)\boldsymbol{\Psi} = \mathbf{0}, \qquad \boldsymbol{\Phi}^H\left(\lambda^2\mathbf{M} + \lambda\mathbf{C} + \mathbf{K}\right) = \mathbf{0}. \tag{1.5}$$

In Eqs. (1.5): $\mathbf{M}^T \neq \mathbf{M}$, $\mathbf{C}^T \neq \mathbf{C}$, $\mathbf{K}^T \neq \mathbf{K}$, $\mathbf{M}, \mathbf{C}, \mathbf{K} \in \mathcal{R}^{N\times N}$ and $\lambda_i \in \mathcal{C}$ ($\lambda_i = \sigma_i + \omega_i\mathrm{i}$, $\lambda_i^* = \sigma_i - \omega_i\mathrm{i}$, $\mathrm{i}^2 = -1$, $\sigma_i, \omega_i \in \mathcal{R}$), while for eigenvectors: $\boldsymbol{\Psi}_i, \boldsymbol{\Phi}_i \in \mathcal{C}^N$ with $i = 1, ..., 2N$.

As can be seen, an asymmetry of the matrices $\mathbf{M}$, $\mathbf{C}$, $\mathbf{K}$ is assumed here – such cases are unusual and rare in the practice of computation and analysis of various technical problems. The asymmetry of the $\mathbf{K}$ matrix appears in the presence of follower forces, the asymmetry of the $\mathbf{M}$

matrix appears in the presence of hydrodynamic forces (i.e., in the flow of fluid around bodies), while the asymmetry of the $\mathbf{C}$ matrix arises in the presence of gyroscopic and/or Coriolis forces.

In the eigenproblem given by Eqs. (1.5), left and right eigenvectors are considered – both the eigenvalues and the corresponding eigenvectors are either complex or real, and, furthermore, the left eigenvector $\mathbf{\Phi}$ and the right eigenvector $\mathbf{\Psi}$ are different. An overview of the formulation QEP with a discussion of its properties can be found in (Tisseur & Meerbergen, 2001; Day & Walsh, 2007). It should be emphasised that Eqs. (1.5) do not represent a classically understood eigenproblem in the sense that the eigenvectors are complex – which causes the vector components to differ in phase with respect to each other. From a physical point of view, this means that standing waves will not appear in the system (as for a classically understood system without damping or with proportional damping), but travelling waves will, because the nodes and antinodes of the vibrational modes do not have a fixed position – they are variable in space and time.

In general, the issue presented in the paper with the non-symmetry of all matrices simultaneously is rarely encountered in engineering practices. In civil engineering problems, asymmetry mainly occurs with follower loads. However, even with symmetric matrices, the sensitivity analysis of multiple eigenvalues presents difficulties because the multiple eigenvalue is always differentiable in the Gâteaux sense, but is not always differentiable in the Fréchet sense.

2. Basic properties of QEP

Equations (1.5) have the so-called trivial (obvious) solutions at $\mathbf{\Psi} = \mathbf{0}$ and $\mathbf{\Phi} = \mathbf{0}$. On the other hand, non-trivial solutions called eigenvectors are obtained when one puts into the λ the solutions of the characteristic equation $W(\lambda)$, i.e., the roots of the polynomial resulting from the expansion of the determinant:

$$W\left(\lambda(\mathbf{h}), \mathbf{h}\right) \equiv \det\left(\mathbf{L}\left(\lambda(\mathbf{h}), \mathbf{h}\right)\right) = 0. \tag{2.1}$$

These solutions (roots) of Eq. (2.1) are called eigenvalues, which can be single (simple) or multiple.

An important property of the characteristic polynomial (Eq. (2.1)) is its differentiability in the sense of Gâteaux at every point $\mathbf{h}$ and in every direction $\mathbf{e}$, while in the Fréchet sense – only beyond the multiple eigenvalues (Balakrishnan, 1976; Tisseur & Meerbergen, 2001; Gekeler, 2008; Seyranian *et al.*, 1994).

The Gâteaux differential of the mapping $\mathbf{g}\left(\mathbf{h}\right) : \mathcal{R}^{N_p} \to \mathcal{R}^n$ at the point $\mathbf{h}$ and in the direction $\mathbf{e}$ is called the mapping $d\widetilde{\mathfrak{G}}\left(\mathbf{h}; \epsilon\mathbf{e}\right)$ such that:

$$\forall_{\epsilon\in\mathcal{R}_+} \quad d\widetilde{\mathfrak{G}}(\mathbf{h};\mathbf{e}) = \lim_{\epsilon\to 0^+} \frac{\mathbf{g}(\mathbf{h}+\epsilon\mathbf{e}) - \mathbf{g}(\mathbf{h})}{\epsilon} \equiv \frac{d}{d\epsilon}\mathbf{g}(\mathbf{h}+\epsilon\mathbf{e})\bigg|_{\epsilon=0} \tag{2.2}$$

under the condition that this limit exists (Gekeler, 2008). If this limit exists, then it (thus the differential $d\widetilde{\mathfrak{G}}$) is determined uniquely. The mapping $d\widetilde{\mathfrak{G}}$ is an element of the space $\mathcal{R}^n$. Equation (2.2) gives a way of calculating the Gâteaux differential – either directly from the limit definition or as the derivative of a function of one variable ϵ. The Gâteaux derivative is always homogeneous

$$\forall_{\alpha\in\mathcal{R}_+} \quad d\widetilde{\mathfrak{G}}(\mathbf{h};\alpha\mathbf{e}) = \alpha d\widetilde{\mathfrak{G}}(\mathbf{h};\mathbf{e}), \tag{2.3}$$

but not always additive, so in general

$$d\widetilde{\mathfrak{G}}(\mathbf{h};\mathbf{e}_1+\mathbf{e}_2) \neq d\widetilde{\mathfrak{G}}(\mathbf{h};\mathbf{e}_1) + d\widetilde{\mathfrak{G}}(\mathbf{h};\mathbf{e}_2). \tag{2.4}$$

The lack of additivity means that the Gâteaux directional derivative is not always a linear operator as the Fréchet operator is.

If the inertia matrix is non-singular $\det(\mathbf{M}) \neq 0$, then the operator $\mathbf{L}$ is regular and the characteristic polynomial has $2N$ finite solutions (Tisseur & Meerbergen, 2001). The set of different roots of the polynomial $W(\lambda)$ is called the spectrum $\widetilde{\mathcal{S}}_{\mathbf{L}}$ of the operator $\mathbf{L}$:

$$\widetilde{\mathcal{S}}_{\mathbf{L}} = \{\lambda_m \in \mathcal{C} \ : \ W(\lambda_m) = 0\}, \qquad m = 1, ..., \widetilde{\Omega}, \tag{2.5}$$

where $\widetilde{\Omega}$ is the number of different roots in the spectrum. Let us mark an important feature of the spectrum $\widetilde{\mathcal{S}}_{\mathbf{L}}$: when the matrices $\mathbf{M}$, $\mathbf{C}$, $\mathbf{K}$ are arbitrary, but real, or are complex, but Hermitian, the spectrum $\widetilde{\mathcal{S}}_{\mathbf{L}}$ is symmetric against the real axis in the complex plane, i.e., the elements of the spectrum appear as real numbers $\lambda \in \mathcal{R}$ or as coupled roots (λ, λ^*).

The concept of the spectrum $\widetilde{\mathcal{S}}_{\mathbf{L}}$, in particular the position of the eigenvalues on the complex plane, plays a fundamental role in the study of mechanical systems, due to the fact that the solution of the eigenproblem is the basis for determining the motion of the system without external forces and excited only by the initial conditions. Such a motion of a mechanical system reveals its inherent properties depending only on the boundary conditions, the distribution of stiffness, the distribution of masses and damping, the susceptibility of the connections, the materials used, the dissipative properties, etc.

In general, among the solutions of Eq. (2.1), there may be real elements $\lambda_i = \sigma_i \pm \mathrm{i}0$ and purely imaginary elements $\lambda_i = 0 \pm \mathrm{i}\omega_i$ and complex elements $\lambda_i = \sigma_i \pm \mathrm{i}\omega_i$. For the assumption of real matrices, the complex elements, if present, are always paired with their conjugate (there is therefore always an even number of them). It is therefore possible (under the assumptions made), without loss of generality, to consider narrowing the spectrum $\widetilde{\mathcal{S}}_{\mathbf{L}}$ to a subset $\mathcal{S}_{\mathbf{L}} \subseteq \widetilde{\mathcal{S}}_{\mathbf{L}}$ containing all real eigenvalues (if existent) and only complex eigenvalues (if existent) with, e.g., negative imaginary parts:

$$\mathcal{S}_{\mathbf{L}} = \Big\{\lambda \in \widetilde{\mathcal{S}}_{\mathbf{L}} \quad : \quad \lambda \in \mathcal{R} \quad \vee \quad (\lambda \in \mathcal{C} \quad \wedge \quad \Im\mathfrak{m}\lambda < 0)\Big\}. \tag{2.6}$$

The number of elements in the spectrum $\mathcal{S}_{\mathbf{L}}$ is denoted by Ω. An additional benefit of the above definition is that a relatively simple way of ordering and numbering the eigenvalues in the spectrum $\mathcal{S}_{\mathbf{L}}$ can be introduced – as will be shown later.

The roots of the characteristic polynomial $W(\lambda)$ can be multiple – the multiplicity of the i-th root λ_i is called its algebraic multiplicity and denoted by $n_a(\lambda_i)$, whereby relation $1 \leq n_a(\lambda_i) \leq 2N$ is true.

Analysing the eigenvectors associated with the eigenvalues, let us point out that Eqs. (1.5) are two different eigenproblems (left and right, respectively) with the same eigenvalues and their algebraic multiplicities (identical spectrum), but in general different complex eigenvectors, each of which has dimensions $[N \times 1]$. The corresponding equations resulting from the conjugations of Eqs. (1.5) are also true:

$$\left(\lambda^{*2}\mathbf{M} + \lambda^*\mathbf{C} + \mathbf{K}\right)\boldsymbol{\Psi}^* = \mathbf{0}, \qquad \left(\lambda^2\mathbf{M}^T + \lambda\mathbf{C}^T + \mathbf{K}^T\right)\boldsymbol{\Phi}^* = \mathbf{0}. \tag{2.7}$$

A comparison of Eqs. (1.5) and (2.7) shows that if the eigenvectors $\boldsymbol{\Psi}$ and $\boldsymbol{\Phi}^*$ are associated with the eigenvalue λ, then the eigenvectors $\boldsymbol{\Psi}^*$ and $\boldsymbol{\Phi}$ are always associated with the conjugate λ^*; the complex eigenvalue λ and its conjugate λ^* may correspond to real eigenvectors, then $\boldsymbol{\Psi} = \boldsymbol{\Psi}^* \in \mathcal{R}^{N\times 1}$ and $\boldsymbol{\Phi} = \boldsymbol{\Phi}^* \in \mathcal{R}^{N\times 1}$ – thus, in this case, a complex pair of eigenvalues corresponds in fact to one right real eigenvector and one left real eigenvector. In contrast, a real eigenvalue always corresponds to real eigenvectors, and it may happen that different real eigenvalues correspond to the same eigenvector. For every eigenvalue, there corresponds at least one left and one right eigenvector; for every left eigenvector, there corresponds a right eigenvector and vice versa. Thus, eigenvectors create separate left and right bases in the eigenspaces.

In the general case, the numbers of eigenvectors resulting from the solution of the eigenproblem (Eqs. (1.5)) may be less than the numbers of eigenvalues together with their multiplicities, i.e., less than $2N$ – in which case we will say that the operator $\mathbf{L}$ is defective, i.e.,

not diagonalizable. The number of linearly independent eigenvectors associated with a given eigenvalue λ_m in the spectrum is called its geometric multiplicity $n_g(\lambda_m)$, the number $n_g(\lambda_m)$ is the same for the left and right eigenvectors. So, there is a fundamental relationship:

$$1 \leq n_g(\lambda_m) \leq n_a(\lambda_m) \leq 2N, \qquad m = 1, ..., \Omega. \tag{2.8}$$

So in particular, it may happen that the number of eigenvectors $n_g(\lambda_m)$ associated with a given eigenvalue λ_m may be less than the algebraic multiplicity $n_a(\lambda_m)$ of this eigenvalue – if $n_g(\lambda_m) < n_a(\lambda_m)$, then the eigenvalue λ_m will be said to be defective; if $n_g(\lambda_m) = n_a(\lambda_m)$, then λ_m is non-defective. In general, the relations between algebraic and geometric multiplicities of eigenvalues are shown in Fig. 1 (Leung, 1993).

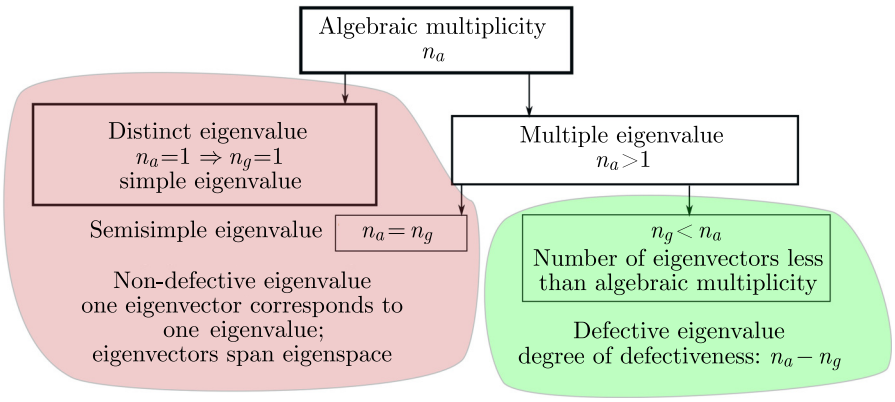

Fig. 1. Relationships between algebraic n_a and geometric n_g multiplicities of eigenvalues and resulting classification (Leung, 1993).

For further analysis, let us write more precisely the eigenproblem corresponding to Eqs. (1.5) with real, non-symmetric matrices $\mathbf{M}$, $\mathbf{C}$, $\mathbf{K}$, which can be represented in two ways. As an eigenproblem with a right eigenvector:

$$\left(\lambda_i^2\mathbf{M} + \lambda_i\mathbf{C} + \mathbf{K}\right)\boldsymbol{\Psi}_i = \mathbf{0}, \qquad i = 1, ..., 2N, \tag{2.9}$$

or with a left eigenvector:

$$\boldsymbol{\Phi}_i^H\left(\lambda_i^2\mathbf{M} + \lambda_i\mathbf{C} + \mathbf{K}\right) = \mathbf{0}, \qquad i = 1, ..., 2N. \tag{2.10}$$

One way to sort the eigenvalues λ_i is the order according to non-decreasing real values. The order introduced here is taken from (Krog & Olhoff, 1995; Olhoff *et al.*, 1995; Seyranian *et al.*, 1994):

$$\underbrace{\underbrace{\widetilde{\lambda}_S, ..., \widetilde{\lambda}_{S_0+1}}_{\Im\mathrm{m}\widetilde{\lambda} < 0}\ \overbrace{\widetilde{\lambda}_{S_0} \leq ... \leq \widetilde{\lambda}_1}^{\widetilde{\sigma}_{i+1} \leq \widetilde{\sigma}_i,\ \Re\mathrm{e}\widetilde{\lambda} = \widetilde{\sigma} < 0}}_{\Im\mathrm{m}\widetilde{\lambda} = \widetilde{\omega} = 0}\ \underbrace{\overbrace{\widehat{\lambda}_1 \leq ... \leq \widehat{\lambda}_{U_0}}^{\widehat{\sigma}_i \leq \widehat{\sigma}_{i+1},\ \Re\mathrm{e}\widehat{\lambda} = \widehat{\sigma} \geq 0}\ \underbrace{\widehat{\lambda}_{U_0+1}, ..., \widehat{\lambda}_U}_{\Im\mathrm{m}\widehat{\lambda} < 0}}_{\Im\mathrm{m}\widehat{\lambda} = \widehat{\omega} = 0}. \tag{2.11}$$

$$\downarrow \widetilde{\lambda}^* \qquad\qquad \downarrow \widehat{\lambda}^*$$

$$\Im\mathrm{m}\widetilde{\lambda}^* > 0 \qquad\qquad \Im\mathrm{m}\widehat{\lambda}^* > 0$$

The specification of the number of solutions is as follows:

$$2N = S_0 + U_0 + 2(S - S_0) + 2(U - U_0). \tag{2.12}$$

In addition, let

$$L = S + U, \tag{2.13}$$

then it can be written:

$$\lambda_i = \begin{cases} \widetilde{\lambda}_{S-i+1} & \text{when } i \leq S, \\ \widehat{\lambda}_{i-S} & \text{when } S < i \leq L, \end{cases} \qquad i = 1, ..., L. \tag{2.14}$$

In particular, it follows from Eq. (2.14) that: $\lambda_1 = \widetilde{\lambda}_S$, $\lambda_{S+1} = \widehat{\lambda}_1$, $\lambda_L = \widehat{\lambda}_U$.

As mentioned, the essence of the ordering thus introduced in Eq. (2.11) is to sort and number all eigenvalues with non-positive imaginary parts (eigenfrequencies) according to the non-decreasing real parts; in addition, the eigenvalues with negative and non-negative real parts $\mathfrak{Re}\lambda$ have been distinguished – $\widetilde{\sigma}$ for $\mathfrak{Re}\lambda < 0$ and $\widehat{\sigma}$ for $\mathfrak{Re}\lambda \geq 0$, respectively. Complex eigenvalues with positive imaginary parts were ordered analogously to their conjugations. The presented ordering of eigenvalues with non-positive imaginary parts refers to the already introduced notion of a spectrum $\mathcal{S}_{\mathbf{L}}$. The elements of this spectrum are the different roots from among those covered by the two upper brackets in Eq. (2.11).

Suppose that there are Ω of different eigenvalues ($\Omega \leq L$) in the spectrum of $\mathcal{S}_{\mathbf{L}}$. Taking into account Eq. (2.14), it is convenient to number them as follows:

$$\check{\lambda}_m = \lambda_i, \qquad i = r_m, r_m + 1, r_m + 2, ..., R_m, \quad m = 1, ..., \Omega, \tag{2.15}$$

where $r_1 = 1$, $r_{m+1} = R_m + 1$, $R_\Omega = L$, and $n_a(\check{\lambda}_m) = R_m - r_m + 1$ is algebraic multiplicity of m-th eigenvalues. Note that with $r_m = R_m$, $n_a(\check{\lambda}_m) = 1$, i.e., $\check{\lambda}_m = \lambda_i$, $i = r_m$ is a single eigenvalue.

An important issue that is considered in terms of eigenvectors is their scaling formula. In eigenvector sensitivity analysis, the formula also plays an important role. The following scaling formula is used in this study:

$$\boldsymbol{\Psi}_j^T \left(2\lambda_j \mathbf{M} + \mathbf{C}\right) \boldsymbol{\Psi}_j = 1, \qquad \boldsymbol{\Phi}_j^T \left(2\lambda_j^* \mathbf{M}^T + \mathbf{C}^T\right) \boldsymbol{\Phi}_j = 1. \tag{2.16}$$

In those formulas, the dependence of the vectors, matrices, and eigenvalues on the parameter $\mathbf{h}$ is not shown for clarity of notation. Let us note here, however, that the proposed method of scaling captures a certain relation of the vector with its transpose, leading to unity, i.e., to a scalar, which is no longer a function of the $\mathbf{h}$, and therefore, its derivative is zero. We will use this scaling property written by Eq. (2.16) as an additional condition in determining the derivatives of the eigenvectors – the derivatives of the mentioned formula will combine the scaled eigenvectors and their derivatives.

For any multiple eigenvalue of $\check{\lambda}_m$, Eqs. (2.9) and (2.10) are true, but they are also true for any linear combinations of eigenvectors corresponding to $\check{\lambda}_m$. This fact can be put as follows for a right eigenvector:

$$\check{\boldsymbol{\Psi}}_j^m = \sum_{k=r_m}^{R_m} \theta_{kj}^m \boldsymbol{\Psi}_k^m, \qquad \theta_{kj}^m \in \mathcal{C}, \quad j = r_m, ..., R_m, \quad m = 1, ..., \Omega, \tag{2.17}$$

and for a left eigenvector:

$$\check{\boldsymbol{\Phi}}_j^m = \sum_{k=r_m}^{R_m} \alpha_{kj}^m \boldsymbol{\Phi}_k^m, \qquad \alpha_{kj}^m \in \mathcal{C}, \quad j = r_m, ..., R_m, \quad m = 1, ..., \Omega, \tag{2.18}$$

where the coefficients θ_{kj}^m and α_{kj}^m form matrices $\boldsymbol{\theta}^m$, $\boldsymbol{\alpha}^m$, respectively. The above two formulas reveal the assumption that the considered multiple eigenvalue θ_{kj}^m is non-defective, because there exist for them $n_g(\theta_{kj}^m) = n_a(\theta_{kj}^m)$ linearly independent eigenvectors. But, as pointed out earlier, determining the basis of the eigenvectors corresponding to a multiple eigenvalue is not even unambiguous as to direction.

Both Eqs. (2.17) and (2.18) can be written for convenience in the respective matrix forms:

$$\begin{aligned} \check{\boldsymbol{\Psi}}^m &= \boldsymbol{\Psi}^m \boldsymbol{\theta}^m, \qquad \boldsymbol{\theta}^m = \left[\theta_{kj}^m\right], \\ \check{\boldsymbol{\Phi}}^m &= \boldsymbol{\Phi}^m \boldsymbol{\alpha}^m, \qquad \boldsymbol{\alpha}^m = \left[\alpha_{kj}^m\right], \end{aligned} \qquad j,k = r_m, ..., R_m, \quad m = 1, ..., \Omega, \tag{2.19}$$

where the matrices of the complex coefficients $\boldsymbol{\theta}^m$ and $\boldsymbol{\alpha}^m$ are quadratic and have the dimension of the algebraic eigenvalue times $\left[n_a(\check{\lambda}_m) \times n_a(\check{\lambda}_m)\right]$; moreover, these matrices are orthogonal, i.e., $\boldsymbol{\theta}^m = \boldsymbol{\theta}^{mH}$, $\boldsymbol{\alpha}^m = \boldsymbol{\alpha}^{mH}$. In contrast, the other matrices

$$\begin{aligned} \boldsymbol{\Psi}^m &= \left[\boldsymbol{\Psi}_{r_m}, ..., \boldsymbol{\Psi}_{R_m}\right], \qquad \check{\boldsymbol{\Psi}}^m = \left[\check{\boldsymbol{\Psi}}_{r_m}^m, ..., \check{\boldsymbol{\Psi}}_{R_m}^m\right], \\ \boldsymbol{\Phi}^m &= \left[\boldsymbol{\Phi}_{r_m}, ..., \boldsymbol{\Phi}_{R_m}\right], \qquad \check{\boldsymbol{\Phi}}^m = \left[\check{\boldsymbol{\Phi}}_{r_m}^m, ..., \check{\boldsymbol{\Phi}}_{R_m}^m\right], \end{aligned} \qquad m = 1, ..., \Omega, \tag{2.20}$$

have dimensions $\left[N \times n_a(\check{\lambda}_m)\right]$. The columns of the matrix $\boldsymbol{\Psi}^m$ are right eigenvectors, and the columns of the matrix $\boldsymbol{\Phi}^m$ are left eigenvectors – these correspond to the multiple eigenvalue $\check{\lambda}_m$ and are the result of solving the eigenproblems (Eqs. (2.9) and (2.10)).

It follows from the assumption that $\check{\lambda}_m$ is a non-defective eigenvalue that the coefficient matrices of $\boldsymbol{\theta}^m$ and $\boldsymbol{\alpha}^m$ are of full rank, i.e:

$$\operatorname{rank}\left(\boldsymbol{\theta}^m\right) = \operatorname{rank}\left(\boldsymbol{\alpha}^m\right) = n_a\left(\check{\lambda}_m\right), \qquad m = 1, ..., \Omega. \tag{2.21}$$

In the remaining part of the text, it will be assumed that the right eigenvectors $\check{\boldsymbol{\Psi}}_j^m$ of Eq. (2.17) and the left eigenvectors $\check{\boldsymbol{\Phi}}_j^m$ of Eq. (2.18), calculated as linear combinations of the corresponding eigenvectors $\boldsymbol{\Psi}_j^m$, $\boldsymbol{\Phi}_j^m$ obtained directly from Eqs. (2.9) and (2.10), were scaled according to Eq. (2.16).

The following equations also remain true for the adopted ordering and designations:

$$\left(\check{\lambda}_m^2 \mathbf{M} + \check{\lambda}_m \mathbf{C} + \mathbf{K}\right) \check{\boldsymbol{\Psi}}_j^m = \mathbf{0}, \qquad j = r_m, ..., R_m, \quad m = 1, ..., \Omega, \tag{2.22}$$

$$\left(\lambda_j^2 \mathbf{M} + \lambda_j \mathbf{C} + \mathbf{K}\right) \check{\boldsymbol{\Psi}}_j^m = \mathbf{0}, \qquad j = r_m, ..., R_m, \quad m = 1, ..., \Omega, \tag{2.23}$$

$$\check{\boldsymbol{\Phi}}_s^{mH} \left(\check{\lambda}_m^2 \mathbf{M} + \check{\lambda}_m \mathbf{C} + \mathbf{K}\right) = \mathbf{0}, \qquad s = r_m, ..., R_m, \quad m = 1, ..., \Omega, \tag{2.24}$$

$$\check{\boldsymbol{\Phi}}_s^{mH} \left(\lambda_s^2 \mathbf{M} + \lambda_s \mathbf{C} + \mathbf{K}\right) = \mathbf{0}, \qquad s = r_m, ..., R_m, \quad m = 1, ..., \Omega. \tag{2.25}$$

3. Directional derivatives of eigenvalues

To compute the directional derivative in the Gâteaux sense, one has to perturb a mapping along the $\mathbf{e}$ direction. In order to find the value of the function at the point $\mathbf{h} + \epsilon\mathbf{e}$, the Taylor expansion around the point $\mathbf{h}$ restricted to the linear part with respect to ϵ is used. Thus, for $m = 1, ..., \Omega$ and $j = r_m, ..., R_m$ the following expansions exists:

$$\lambda_j\left(\mathbf{h} + \epsilon\mathbf{e}\right) = \check{\lambda}_m\left(\mathbf{h}\right) + \epsilon\mu_j, \qquad \lambda_j^*\left(\mathbf{h} + \epsilon\mathbf{e}\right) = \check{\lambda}_m^*\left(\mathbf{h}\right) + \epsilon\mu_j^*, \tag{3.1}$$

$$\check{\boldsymbol{\Psi}}_j^m\left(\mathbf{h} + \epsilon\mathbf{e}\right) = \check{\boldsymbol{\Psi}}_j^m\left(\mathbf{h}\right) + \epsilon\boldsymbol{\Gamma}_j, \qquad \check{\boldsymbol{\Phi}}_j^m\left(\mathbf{h} + \epsilon\mathbf{e}\right) = \check{\boldsymbol{\Phi}}_j^m\left(\mathbf{h}\right) + \epsilon\boldsymbol{\Pi}_j, \tag{3.2}$$

where $\mu_j \in \mathcal{C}$, $\boldsymbol{\Gamma}_j = [\Gamma_{rj}]$ and $\Gamma_{rj} \in \mathcal{C}$, $\boldsymbol{\Pi}_j = [\Pi_{rj}]$ and $\Pi_{rj} \in \mathcal{C}$, $r = 1, ..., N$ are the sought directional derivatives of the j-th eigenvalue and the associated right and left directional derivatives of the eigenvector, respectively; it follows from the above notations that the derivatives mentioned here are complex, so they consist of derivatives of real parts and derivatives of imaginary parts.

The basis for the derivation of directional derivatives in the Gâteaux sense of the eigenvalues of Eq. (2.23) and (2.22) is to write them at the point $\mathbf{h} + \epsilon\mathbf{e}$ for $j = r_m, ..., R_m$, $m = 1, ..., \Omega$:

$$\left[\check{\lambda}_m^2(\mathbf{h} + \epsilon\mathbf{e})\mathbf{M}(\mathbf{h} + \epsilon\mathbf{e}) + \check{\lambda}_m(\mathbf{h} + \epsilon\mathbf{e})\mathbf{C}(\mathbf{h} + \epsilon\mathbf{e}) + \mathbf{K}(\mathbf{h} + \epsilon\mathbf{e})\right] \check{\boldsymbol{\Psi}}_j^m(\mathbf{h} + \epsilon\mathbf{e}) = \mathbf{0}. \tag{3.3}$$

By substituting Eqs. (3.1), (3.2), (1.4) into Eq. (3.3) one obtains for $j = r_m, ..., R_m$, $m = 1, ..., \Omega$:

$$\left[\left(\check{\lambda}_m^2 + 2\epsilon \check{\lambda}_m \mu_j + \epsilon^2 \mu_j^2 \right) \left(\mathbf{M} + \epsilon \sum_{p=1}^{N_p} \frac{\partial \mathbf{M}}{\partial h_p} e_p \right) \right.$$
$$\left. + \left(\check{\lambda}_m + \epsilon \mu_j \right) \left(\mathbf{C} + + \epsilon \sum_{p=1}^{N_p} \frac{\partial \mathbf{C}}{\partial h_p} e_p \right) + \mathbf{K} + \epsilon \sum_{p=1}^{N_p} \frac{\partial \mathbf{K}}{\partial h_p} e_p \right] \left(\check{\mathbf{\Psi}}_j^m + \epsilon \mathbf{\Gamma}_j \right) = \mathbf{0}. \tag{3.4}$$

Further transformations of Eq. (3.4) take into account that it is true for any $\epsilon > 0$, non-linear terms relative to ϵ are omitted, and that Eq. (3.2) is true, and also Eq. (2.23) is substituted. After simplifying, the sums give the following equations:

$$\left(\check{\lambda}_m^2 \mathbf{M} + \check{\lambda}_m \mathbf{C} + \mathbf{K} \right) \mathbf{\Gamma}_j + \left(\check{\lambda}_m^2 \sum_{p=1}^{N_p} \frac{\partial \mathbf{M}}{\partial h_p} e_p + \check{\lambda}_m \sum_{p=1}^{N_p} \frac{\partial \mathbf{C}}{\partial h_p} e_p + \sum_{p=1}^{N_p} \frac{\partial \mathbf{K}}{\partial h_p} e_p \right) \check{\mathbf{\Psi}}_j^m$$
$$+ \mu_j \left(2\check{\lambda}_m \mathbf{M} + \mathbf{C} \right) \check{\mathbf{\Psi}}_j^m = \mathbf{0}, \qquad j = r_m, ..., R_m, \quad m = 1, ..., \Omega. \tag{3.5}$$

Before discussing further, let us introduce for convenience the following designations for $m = 1, ..., \Omega$:

$$\widehat{\mathbf{E}}^m = \check{\lambda}_m^2 \mathbf{M} + \check{\lambda}_m \mathbf{C} + \mathbf{K}, \qquad \widehat{\mathbf{G}}^m = 2\check{\lambda}_m \mathbf{M} + \mathbf{C},$$
$$\widehat{\mathbf{F}}^m = \sum_{p=1}^{N_p} \left(\check{\lambda}_m^2 \frac{\partial \mathbf{M}}{\partial h_p} + \check{\lambda}_m \frac{\partial \mathbf{C}}{\partial h_p} + \frac{\partial \mathbf{K}}{\partial h_p} \right) e_p. \tag{3.6}$$

Taking into account the above, Eq. (3.5) can be written in the form:

$$\widehat{\mathbf{E}}^m \mathbf{\Gamma}_j + \widehat{\mathbf{F}}^m \check{\mathbf{\Psi}}_j^m + \widehat{\mathbf{G}}^m \check{\mathbf{\Psi}}_j^m \mu_j = \mathbf{0}, \qquad j = r_m, ..., R_m, \quad m = 1, ..., \Omega. \tag{3.7}$$

Multiplying Eq. (3.7) left-hand by $\mathbf{\Phi}_s^H$, we notice that the first component of $\mathbf{\Phi}_s^H \widehat{\mathbf{E}}^m$ disappears, because there is Eq. (2.24), and Eq. (2.17) should also be considered here. Introducing the simplified notations $\mathbf{A}^m = [a_{sk}^m]$, $\mathbf{B}^m = [b_{sk}^m]$ where

$$a_{sk}^m = \mathbf{\Phi}_s^H \widehat{\mathbf{F}}^m \mathbf{\Psi}_k, \qquad b_{sk}^m = \mathbf{\Phi}_s^H \widehat{\mathbf{G}}^m \mathbf{\Psi}_k, \tag{3.8}$$

finally, Eq. (3.5) is obtained in the form:

$$\sum_{k=r_m}^{R_m} \left(a_{sk}^m + \mu_j b_{sk}^m \right) \theta_{kj}^m = 0, \qquad j, s = r_m, ..., R_m, \quad m = 1, ..., \Omega. \tag{3.9}$$

The following additional eigenproblem arises from Eq. (3.9):

$$\left(\mathbf{A}^m + \mu_j \mathbf{B}^m \right) \boldsymbol{\theta}_j^m = \mathbf{0}, \qquad j = r_m, ..., R_m, \quad m = 1, ..., \Omega. \tag{3.10}$$

The matrices $\mathbf{A}^m$ and $\mathbf{B}^m$ are, in general, asymmetric and complex, and $\boldsymbol{\theta}_j^m = \left[\theta_{kj}^m \right]$. From Eq. (3.10), one computes $n_a(\check{\lambda}_m)$ of the eigenvalues, i.e., derivatives of $\mu_{r_m}, ..., \mu_{R_m}$, which are being sought. An important observation is that if the matrices $\mathbf{A}^m$ and $\mathbf{B}^m$ are both diagonal (off-diagonal terms are equal to zero), the eigenvalues μ_j, i.e., Gatêaux directional derivatives of multiple eigenvalues $\check{\lambda}_m$ are the same as the traditional Fréchet derivatives. The diagonal form

of matrices $\mathbf{A}^m$ and $\mathbf{B}^m$ occurs when matrices $\mathbf{F}^m$ and $\mathbf{G}^m$ from Eq. (3.6) are simultaneously diagonalizable by transformations (Eq. (3.8)). Of course, for a single eigenvalue $n_a(\check{\lambda}_m) = 1$ (scalar equation), these derivatives are also the same.

Equation (3.10) has, in practice, a low dimension associated with the algebraic multiplicities of the multiple eigenvalue $\check{\lambda}_m$. However, for a single (simple) eigenvalue, when, for a given m, the algebraic multiplicity $n_a(\check{\lambda}_m) = 1$ (then $j = k = s = r_m$), Eq. (3.10) turns into a scalar equation from which the directional derivative μ_j of the simple eigenvalue λ_j is calculated. Thus, the derived relations leading to (Eq. (3.10)) constitute an algorithm for calculating the directional derivatives of the eigenvalues of the eigenproblem (Eq. (2.22)) regardless of their multiplicity.

4. Directional derivatives of eigenvectors

The derivation of the derivative of the right eigenvector is shown here – this may be done analogously for the left eigenvector. The source is Eq. (3.7), because in it the derivative of the eigenvector $\mathbf{\Gamma}_j$ appears in the process of differentiation. Let us rewrite Eq. (3.7) in a more convenient form:

$$\widehat{\mathbf{E}}^m\mathbf{\Gamma}_j = -\widehat{\mathbf{F}}^m\check{\mathbf{\Psi}}_j^m - \widehat{\mathbf{G}}^m\check{\mathbf{\Psi}}_j^m\mu_j, \qquad j = r_m, ..., R_m, \quad m = 1, ..., \Omega. \tag{4.1}$$

If the derivatives of the eigenvalues μ_j are already known, as they are calculated from Eq. (3.10), the right-hand side of Eq. (4.1) is different from zero and it is theoretically possible to determine $\mathbf{\Gamma}_j$, i.e., the derivatives of the right eigenvector.

However, the fundamental difficulty in Eq. (4.1) relates to the fact that, since $\check{\lambda}_m$ is an eigenvalue with multiplicity $n_a(\check{\lambda}_m)$, so the matrix $\mathbf{E}^m$ is singular (rank $\mathbf{E}^m$ is reduced by $n_a(\check{\lambda}_m)$) – it is therefore not possible to simply compute $\mathbf{\Gamma}_j$. Overcoming this difficulty has been the subject of numerous research papers and related computational algorithms of varying complexity and difficulty of application, both in configuration space and state space. An extensive review of the papers in this area is given in (Choi *et al.*, 2004).

The algorithms presented in (Kim *et al.*, 1999a; 1999b; Choi *et al.*, 2004; Lee *et al.*, 1999a; 1999b) are used in the paper but extended to non-symmetric matrices in QEP. The idea is to use the eigenvector scaling Eq. (2.16) for the right eigenvector as a constraint condition – because, as mentioned, this equation, when differentiated, ties the vector $\mathbf{\Psi}$ itself to its derivative $\mathbf{\Gamma}_j$. The scaling condition is written here in the formula for multiple eigenvalues:

$$\check{\mathbf{\Psi}}_j^{mT}(\mathbf{h})\left[2\check{\lambda}_m(\mathbf{h})\mathbf{M}(\mathbf{h}) + \mathbf{C}(\mathbf{h})\right]\check{\mathbf{\Psi}}_j^m(\mathbf{h}) = 1, \qquad j = r_m, ..., R_m, \quad m = 1, ..., \Omega. \tag{4.2}$$

Thus, let us expand the scaling condition (Eq. (4.2)) at the point $\mathbf{h}$ and in the direction $\mathbf{e}$ (analogous to Eq. (3.3)), however, under the assumption that the vectors $\check{\mathbf{\Psi}}_j^m$ are already scaled according to Eq. (4.2), (the scaling should therefore be done after solving Eq. (3.10) and performing a linear combination defined in Eq. (2.19)), so for $j = r_m, ..., R_m,\ m = 1, ..., \Omega$:

$$\check{\mathbf{\Psi}}_j^{mT}(\mathbf{h}+\epsilon\mathbf{e})\left[2\check{\lambda}_m(\mathbf{h}+\epsilon\mathbf{e})\mathbf{M}(\mathbf{h}+\epsilon\mathbf{e}) + \mathbf{C}(\mathbf{h}+\epsilon\mathbf{e})\right]\check{\mathbf{\Psi}}_j^m(\mathbf{h}+\epsilon\mathbf{e}) = 1. \tag{4.3}$$

Substituting Eqs. (3.1), (3.2), (1.4) to Eq. (4.3), the following is obtained:

$$\left(\check{\mathbf{\Psi}}_j^{mT} + \epsilon\mathbf{\Gamma}_j^T\right)\left[2\left(\check{\lambda}_m + \epsilon\mu_j\right)\left(\mathbf{M} + \epsilon\sum_{p=1}^{N_p}\frac{\partial\mathbf{M}}{\partial h_p}e_p\right) + \mathbf{C} + \epsilon\sum_{p=1}^{N_p}\frac{\partial\mathbf{C}}{\partial h_p}e_p\right]\left(\check{\mathbf{\Psi}}_j^m + \epsilon\mathbf{\Gamma}_j\right) = 1. \tag{4.4}$$

Further transformations of Eq. (4.4) take into account that it is true for any $\epsilon > 0$, nonlinear members with respect to ϵ are omitted, and that Eq. (3.2) is considered, also Eqs. (2.23), (2.16) are substituted. The final expression after ordering the sums is obtained in the form:

$$\breve{\boldsymbol{\Psi}}_j^{mT}\left[2\breve{\lambda}_m\left(\mathbf{M}+\mathbf{M}^T\right)+\mathbf{C}+\mathbf{C}^T\right]\boldsymbol{\Gamma}_j+2\breve{\boldsymbol{\Psi}}_j^{mT}\mathbf{M}\breve{\boldsymbol{\Psi}}_j^m\mu_j$$
$$=-\breve{\boldsymbol{\Psi}}_j^{mT}\left(2\breve{\lambda}_m\sum_{p=1}^{N_p}\frac{\partial\mathbf{M}}{\partial h_p}e_p+\sum_{p=1}^{N_p}\frac{\partial\mathbf{C}}{\partial h_p}e_p\right)\breve{\boldsymbol{\Psi}}_j^m. \tag{4.5}$$

For clarity let us introduce the notations:

$$\widehat{\mathbf{H}}^m=\widehat{\mathbf{G}}^{mT}+\widehat{\mathbf{G}}^m,\qquad \widehat{\mathbf{X}}^m=\sum_{p=1}^{N_p}\left(2\breve{\lambda}_m\frac{\partial\mathbf{M}}{\partial h_p}+\frac{\partial\mathbf{C}}{\partial h_p}\right)e_p,\quad m=1,...,\Omega. \tag{4.6}$$

Finally, Eq. (4.5) takes a simpler form:

$$\breve{\boldsymbol{\Psi}}_j^{mT}\widehat{\mathbf{H}}^m\boldsymbol{\Gamma}_j+2\breve{\boldsymbol{\Psi}}_j^{mT}\mathbf{M}\breve{\boldsymbol{\Psi}}_j^m\mu_j=-\breve{\boldsymbol{\Psi}}_j^{mT}\widehat{\mathbf{X}}^m\breve{\boldsymbol{\Psi}}_j^m,\qquad j=r_m,...,R_m,\quad m=1,...,\Omega. \tag{4.7}$$

Note that Eqs. (4.1) and (4.7) tie the derivatives of the eigenvalue of μ_j and the derivatives of the eigenvectors of $\boldsymbol{\Gamma}_j$ – so let us write the two equations together, finally for $j=r_m,...,R_m$, $m=1,...,\Omega$ we have

$$\begin{cases}\widehat{\mathbf{E}}^m\boldsymbol{\Gamma}_j+\widehat{\mathbf{G}}^m\breve{\boldsymbol{\Psi}}_j^m\mu_j=-\widehat{\mathbf{F}}^m\breve{\boldsymbol{\Psi}}_j^m,\\ \breve{\boldsymbol{\Psi}}_j^{mT}\widehat{\mathbf{H}}^m\boldsymbol{\Gamma}_j+2\breve{\boldsymbol{\Psi}}_j^{mT}\mathbf{M}\breve{\boldsymbol{\Psi}}_j^m\mu_j=-\breve{\boldsymbol{\Psi}}_j^{mT}\widehat{\mathbf{X}}^m\breve{\boldsymbol{\Psi}}_j^m.\end{cases} \tag{4.8}$$

The above system of equations is still unsolvable for a multiple eigenvalue, i.e., with $n_a\left(\breve{\lambda}_m\right)>1$ – the second equation is a scalar equation and, as noted earlier, the rank of the matrix $\widehat{\mathbf{E}}^m$ is equal:

$$\operatorname{rank}\left(\widehat{\mathbf{E}}^m\right)=N-n_a\left(\breve{\lambda}_m\right).$$

Let us introduce the following diagonal auxiliary matrices:

$$\boldsymbol{\mu}^m=\begin{bmatrix}\mu_{r_m}&0&\cdots&0\\0&\mu_{r_m+1}&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&\mu_{R_m}\end{bmatrix},\qquad \mathbf{Z}^m=2\begin{bmatrix}\widehat{Z}_{r_m}^m&0&\cdots&0\\0&\widehat{Z}_{r_m+1}^m&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&\widehat{Z}_{R_m}^m\end{bmatrix},$$
$$\mathbf{V}^m=\begin{bmatrix}\widehat{V}_{r_m}^m&0&\cdots&0\\0&\widehat{V}_{r_m+1}^m&\cdots&0\\\vdots&\vdots&\ddots&\vdots\\0&0&\cdots&\widehat{V}_{R_m}^m\end{bmatrix},\qquad \widehat{Z}_j^m=\breve{\boldsymbol{\Psi}}_j^{mT}\mathbf{M}\breve{\boldsymbol{\Psi}}_j^m,\qquad \widehat{V}_j^m=\breve{\boldsymbol{\Psi}}_j^{mT}\widehat{\mathbf{X}}^m\breve{\boldsymbol{\Psi}}_j^m, \tag{4.9}$$

and the derivative matrix:

$$\boldsymbol{\Gamma}^m=\left[\boldsymbol{\Gamma}_{r_m}\ \ \boldsymbol{\Gamma}_{r_m+1}\ \ \cdots\ \ \boldsymbol{\Gamma}_{R_m}\right]. \tag{4.10}$$

The dimensions of the predefined matrices are as follows:

$$\widehat{\mathbf{E}}^m\longrightarrow[N\times N],\qquad \widehat{\mathbf{G}}^m\breve{\boldsymbol{\Psi}}^m\longrightarrow\left[N\times n_a\left(\breve{\lambda}_m\right)\right],$$
$$\breve{\boldsymbol{\Psi}}^{mT}\widehat{\mathbf{H}}^m\longrightarrow\left[n_a\left(\breve{\lambda}_m\right)\times N\right],\qquad \mathbf{Z}^m\longrightarrow\left[n_a\left(\breve{\lambda}_m\right)\times n_a\left(\breve{\lambda}_m\right)\right],$$
$$\boldsymbol{\Gamma}^m\longrightarrow\left[N\times n_a\left(\breve{\lambda}_m\right)\right],\qquad \boldsymbol{\mu}^m\longrightarrow\left[n_a\left(\breve{\lambda}_m\right)\times n_a\left(\breve{\lambda}_m\right)\right],$$
$$\widehat{\mathbf{F}}^m\breve{\boldsymbol{\Psi}}^m\longrightarrow\left[N\times n_a\left(\breve{\lambda}_m\right)\right],\qquad \mathbf{V}^m\longrightarrow\left[n_a\left(\breve{\lambda}_m\right)\times n_a\left(\breve{\lambda}_m\right)\right].$$

Note that the rank of the individual submatrices in Eq. (4.11) are as follows:

$$\operatorname{rank}\left(\check{\boldsymbol{\Psi}}^m\right) = \operatorname{rank}\left(\check{\boldsymbol{\Psi}}^{mT}\right) = n_a\left(\check{\lambda}_m\right),$$

$$\operatorname{rank}\left(\mathbf{Z}^m\right) = \operatorname{rank}\left(\widehat{\mathbf{G}}^m\check{\boldsymbol{\Psi}}^m\right) = \operatorname{rank}\left(\check{\boldsymbol{\Psi}}^{mT}\widehat{\mathbf{H}}^m\right) = n_a\left(\check{\lambda}_m\right).$$

Using these definitions, the following matrix equation can be written in the form:

$$\underbrace{\begin{bmatrix} \widehat{\mathbf{E}}^m & \widehat{\mathbf{G}}^m\check{\boldsymbol{\Psi}}^m \\ \check{\boldsymbol{\Psi}}^{mT}\widehat{\mathbf{H}}^m & \mathbf{Z}^m \end{bmatrix}}_{\mathbf{S}^m} \underbrace{\begin{bmatrix} \boldsymbol{\Gamma}^m \\ \boldsymbol{\mu}^m \end{bmatrix}}_{\mathbf{Y}^m} = \underbrace{-\begin{bmatrix} \widehat{\mathbf{F}}^m\check{\boldsymbol{\Psi}}^m \\ \mathbf{V}^m \end{bmatrix}}_{\mathbf{R}^m}. \tag{4.11}$$

Taking into account the previously listed dimensions of the component matrices in Eq. (4.11), the following relations occur:

$$\mathbf{S}^m \longrightarrow \left[\left(N + n_a\left(\check{\lambda}_m\right)\right) \times \left(N + n_a\left(\check{\lambda}_m\right)\right)\right], \qquad \mathbf{Y}^m, \mathbf{R}^m \longrightarrow \left[\left(N + n_a\left(\check{\lambda}_m\right)\right) \times n_a\left(\check{\lambda}_m\right)\right].$$

The key observation here is that the quadratic and complex coefficient matrix $\mathbf{S}^m$ is not singular (the method of proof is shown in (Kim *et al.*, 1999b)), i.e., it is invertible and thus, there is an unambiguous solution to the system (Eq. (4.11)) – in this equation, the unknown is the matrix $\mathbf{Y}^m$; moreover, the system of algebraic Eq. (4.11) is well-conditioned. The algorithm presented here leads to the simultaneous determination of the directional derivatives of the right eigenvectors $\boldsymbol{\Gamma}_j$ and the redetermination of the directional derivative μ_j associated with the eigenvalue of λ_j by solving a linear algebraic non-symmetric system of complex Eq. (4.11). For a single eigenvalue, when for a given m the algebraic multiplicity $n_a(\check{\lambda}_m) = 1$, the system of Eq. (4.11) also allows us to calculate the desired quantities. Thus, the derived relations constitute an algorithm for computing the directional derivatives of the eigenvectors of Eq. (2.22), irrespective of the eigenvalue multiplicity.

The structure of Eq. (4.11) also indicates that the vector of the directional derivative of the eigenvector $\boldsymbol{\Gamma}_j$ is determined for the same multiplier as the corresponding eigenvector of $\check{\boldsymbol{\Psi}}_j^m$.

5. Additional requirements

It is important to pick up the significant practical problems associated with the determination and recognition of multiple eigenvalues of λ_m (see, e.g., (Seyranian *et al.*, 1994)). Numerical procedures iterate over the eigenvalues of λ_i and the corresponding eigenvectors $\boldsymbol{\Psi}_i$, $\boldsymbol{\Phi}_i$ with some accuracy. Most often, only the tolerance of the $\epsilon_\lambda^{\text{tol}} = 1 \times 10^{-5} \div 1 \times 10^{-3}$ eigenvalue determination is established. Structural models, created at the design stage and characterized by symmetries, may exhibit the presence of multiple eigenvalues, or there may be very closely lying single eigenvalues. This raises the problem of identifying multiple eigenvalues in the spectrum. Thus, an additional algorithm for the recognition of multiple eigenvalues after ordering and renumbering the eigenvalues according to Eq. (2.11) must be applied to the derivation formulas presented.

The simplest way to do this is to take the number ϵ_m^{tol} such that $\epsilon_\lambda^{\text{tol}} < \epsilon_m^{\text{tol}} \ll 1$ as the tolerance for recognising multiple eigenvalues. Thus, referring to the numbering of Eq. (2.14) we will say that two eigenvalues are multiple when:

$$|\lambda_i - \lambda_{i+1}| \le \epsilon_m^{\text{tol}}, \tag{5.1}$$

which implies that they are at a distance less than ϵ_m^{tol} on the complex plane $\sigma - \omega$. Another proposal is to adopt separate distance conditions for eigendamping and eigenfrequency:

$$|\mathfrak{Re}\lambda_i - \mathfrak{Re}\lambda_{i+1}| \le \epsilon_m^{\text{tol}_\sigma} \qquad \text{and} \qquad |\mathfrak{Im}\lambda_i - \mathfrak{Im}\lambda_{i+1}| \le \epsilon_m^{\text{tol}_\omega}, \tag{5.2}$$

which means that two eigenvalues are multiple if they lie in the rectangle $\epsilon_m^{\mathrm{tol}_\sigma} \times \epsilon_m^{\mathrm{tol}_\omega}$ on the complex plane $\sigma - \omega$.

It should be noted that, in general cases, the definitions given above do not allow explicit grouping of multiple eigenvalues, but they do allow such grouping in many analyses of typical structures with clearly separated multiple eigenvalues in the spectrum. Unfortunately, the values of tolerances $\epsilon_m^{\mathrm{tol}}$, $\epsilon_m^{\mathrm{tol}_\sigma}$, $\epsilon_m^{\mathrm{tol}_\omega}$ may be problem dependent.

6. Numerical examples

6.1. Introduction

As an example of a system with multiple eigenvalues, we will present the planar frame shown in Fig. 2. The geometric dimensions are as follows: beam span $L = 16.16855\,\mathrm{m}$, height of columns $H = 5\,\mathrm{m}$. All cross-sections are square in shape. The following parameters were distinguished for further analysis: width of cross-sections $b = 0.4\,\mathrm{m}$, height of column cross-sections $h_s = 0.4\,\mathrm{m}$, height of beam cross-sections $h_r = 0.4\,\mathrm{m}$. In addition, two local zones were distinguished: 1) the right frame node – the cross-section height of these elements $h_n = 0.4\,\mathrm{m}$, and 2) the right column fixed – the cross-section height $h_f = 0.4\,\mathrm{m}$. All mentioned zones span over along one finite beam element.

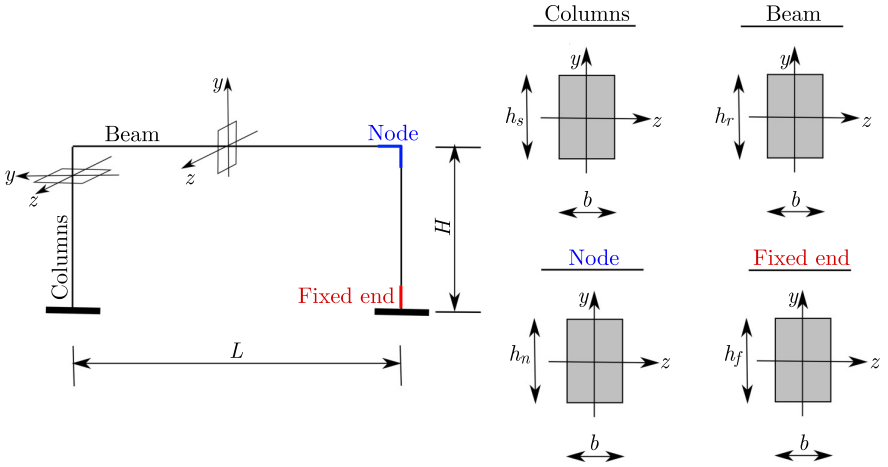

Fig. 2. Flat symmetric frame with fixed columns – geometric and cross-section dimensions.

It was assumed that the frame was made of reinforced concrete: Young's modulus of the material equal to $E = 30\,\mathrm{GPa}$, Poisson's ratio $\nu = 0.2$ and a density $\rho = 2400\,\mathrm{kg/m^3}$. In addition, for simplicity, a damping matrix according to Rayleigh was assumed, i.e., $\mathbf{C} = \alpha\mathbf{M} + \beta\mathbf{K}$, where $\alpha = 2 \times 10^{-3}$, $\beta = 5.2 \times 10^{-4}$.

The geometric dimensions and cross-sections are set so that the first eigenvalue of λ_1 corresponding to the antisymmetric mode shape is equal to λ_2 corresponding to the symmetric mode shape, i.e., $\lambda_1 = \lambda_2$ – Fig. 3. Finally, the symmetric mode shapes correspond to λ_2 and λ_4, while the antisymmetric mode shapes correspond to λ_1 and λ_3.

In the analysed frame, the matrices of the system $\mathbf{M}$, $\mathbf{C}$, $\mathbf{K}$ are symmetric and the eigenmodes are real (classical) – which makes them easier to visualize and in this case the right and left eigenvectors are equal $\mathbf{\Phi} = \mathbf{\Psi} \in \mathcal{R}^N$. The assumptions made do not change the fact that the multiple eigenvalue $\lambda_1 = \lambda_2$ (i.e., $n_a(\lambda_1) = 2$) is differentiable in the Gâteaux sense but is not always differentiable in the Fréchet sense. The tolerance for recognizing multiple eigenvalues in Eq. (5.1) is set to $\epsilon_m^{\mathrm{tol}} = 1 \times 10^{-5}$.

To solve the derivatives shown, a programme in MATLAB was developed to calculate the directional derivatives of the eigenvalues and eigenvectors. The procedure for calculating eigenvalues and eigenvectors for QEP shown in (Hammarling *et al.*, 2013) was used.

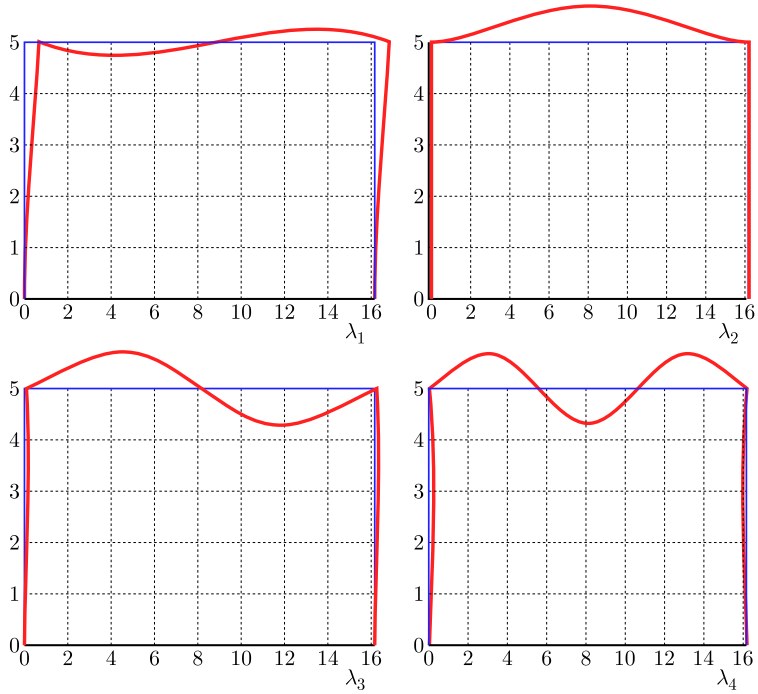

Fig. 3. Eigenmode shapes for consecutive eigenvalues.

6.2. Numerical example 1: Symmetric changes of the frame

Let the vector of parameters against which the derivatives are calculated be equal:

$$\mathbf{h} = [L, H, b, h_s, h_r], \qquad \mathbf{h} \in \mathcal{R}^5. \tag{6.1}$$

Thus, the dimension of parameter space is equal to $N_p = 5$. Table 1 shows directional derivatives with respect to canonical basis vectors (for example, the basis vector $\mathbf{e}_3 = [0, 0, 1, 0, 0]$). It is noteworthy that a change in any component of the vector $\mathbf{h}$ results in a symmetric change in the geometry and cross sections of the frame members. In such a case, the Gâteaux directional derivatives are the same as the Fréchet derivatives even for multiple eigenvalues, i.e.:

$$\frac{\partial \lambda_i}{\partial h_p} \equiv \mu_i(\mathbf{h}, \mathbf{e}_p), \qquad i = 1, ..., 4, \qquad p = 1, ..., 5.$$

The reason for this property is that the matrices $\mathbf{A}^m$ and $\mathbf{B}^m$ are both diagonal in Eq. (3.10). The diagonal form of matrices $\mathbf{A}^m$ and $\mathbf{B}^m$ occurs because matrices $\mathbf{F}^m$ and $\mathbf{G}^m$ from Eq. (3.6) are

Table 1. Eigenvalues and their directional derivatives along the vectors of canonical basis $\mathbf{e}_p$ of the parameter space $\mathbf{h} = [L, H, b, h_s, h_r]$.

λ_i	Deflected shape	$\sigma_i + \omega_i$	$\frac{\partial \lambda_i}{\partial L} \equiv \mu_i(\mathbf{h}, \mathbf{e}_1)$	$\frac{\partial \lambda_i}{\partial H} \equiv \mu_i(\mathbf{h}, \mathbf{e}_2)$	$\frac{\partial \lambda_i}{\partial b} \equiv \mu_i(\mathbf{h}, \mathbf{e}_3)$	$\frac{\partial \lambda_i}{\partial h_s} \equiv \mu_i(\mathbf{h}, \mathbf{e}_4)$	$\frac{\partial \lambda_i}{\partial h_r} \equiv \mu_i(\mathbf{h}, \mathbf{e}_5)$
λ_1	antisym frame	−0.20516 +28.0212i	+0.01665 −1.14285i	+0.10935 −7.50359i	−1.01922 +69.9413i	−0.26076 +17.8940i	+0.26076 −17.8940i
λ_2	sym beam	−0.20516 +28.0212i	+0.04644 −3.18711i	+0.01275 −0.87524i	−1.01598 +69.7185i	−0.15744 +10.8039i	+0.15744 −10.8039i
λ_3	antisym beam	−1.69186 +80.6252i	+0.37242 −8.87510i	+0.14323 −3.41343i	−8.38976 +199.937i	−0.97417 +23.2155i	+0.97417 −23.2155i
λ_4	sym beam	−6.37071 +156.392i	+1.41240 −17.3101i	+0.49267 −6.03818i	−31.4011 +384.846i	−2.37618 +29.1220i	+2.37618 −29.1220i

simultaneously diagonalizable by transformations (Eq. (3.8)), and to explain it deeply, this is because Fréchet derivatives of system matrices ($\partial\mathbf{M}/\partial h_p$, $\partial\mathbf{C}/\partial h_p$, $\partial\mathbf{K}/\partial h_p$ – see Eq. (3.6)) are simultaneously diagonalizable.

6.3. Numerical example 2: Asymmetrical changes of the frame

In contrast to the previous example, let the vector of parameters against which the derivatives are calculated be equal:

$$\mathbf{h} = [h_n, h_f], \qquad \mathbf{h} \in \mathcal{R}^2. \tag{6.2}$$

Thus, the dimension of parameter space is equal to $N_p = 2$. The canonical basis vectors are equal to $\mathbf{e}_1 = [1, 0]$ and $\mathbf{e}_2 = [0, 1]$. It is worth noting that, in this example, a change in any component of the vector $\mathbf{h}$ results in an asymmetric change in the height of the cross-sections of the frame bars. In such a case, the directional derivatives of Gâteaux are not the same as the Fréchet derivatives for a multiple eigenvalue, i.e., there is

$$\frac{\partial\lambda_i}{\partial h_p} \neq \mu_i(\mathbf{h}, \mathbf{e}_p), \qquad i = 1, 2, \quad p = 1, 2.$$

On the other hand, for the simple eigenvalues, the Gâteaux directional derivatives are equal to the Fréchet derivatives, i.e.:

$$\frac{\partial\lambda_i}{\partial h_p} \equiv \mu_i(\mathbf{h}, \mathbf{e}_p), \qquad i = 3, 4, \quad p = 1, 2.$$

Table 2 shows the directional derivatives with respect to the canonical basis vectors of the double eigenvalue of $\lambda_1 = \lambda_2$. The third column of Table 2 shows that in this case the additivity condition for the directional derivative of Gâteaux does not hold – which excludes the existence of Fréchet derivative at the point $\mathbf{h}$. The derivative matrix $\widehat{\mathbf{F}}^m$ is not diagonalizable by transformation $\boldsymbol{\Psi}_s^T\widehat{\mathbf{F}}^m\boldsymbol{\Psi}_k$, although matrix $\widehat{\mathbf{G}}^m$ is diagonalizable by transformation $\boldsymbol{\Psi}_s^T\widehat{\mathbf{G}}^m\boldsymbol{\Psi}_k$ – see Eqs. (3.6) and (3.8).

Table 2. Eigenvalues $\lambda_1 = \lambda_2$ and their directional derivatives along the vectors of canonical basis $\mathbf{e}_p$ of the parameter space $\mathbf{h} = [h_n, h_f]$.

λ_i	Deflected shape	$\sigma_i + \omega_i$	$\mu_i(\mathbf{h}, \mathbf{e}_1)$	$\mu_i(\mathbf{h}, \mathbf{e}_2)$	$\mu_i(\mathbf{h}, \mathbf{e}_1 + \mathbf{e}_2) \neq \mu_i(\mathbf{h}, \mathbf{e}_1) + \mu_i(\mathbf{h}, \mathbf{e}_2)$
λ_1	antisym frame	−0.20516 +28.0212i	+0.01177 −0.80738i	−0.02963 +2.03316i	+0.00007 −0.00474i
λ_2	sym beam	−0.20516 +28.0212i	−0.04451 +3.05466i	−0.00002 +0.00124i	−0.06246 +4.28642i

Table 3 shows the directional derivatives with respect to the canonical basis vectors of the simple eigenvalues λ_3 and λ_4. In this case, the derivative along an arbitrary direction can be calculated from the scalar product of the gradient vector and the direction vector – as for the traditional Fréchet derivative.

Table 3. Eigenvalues λ_3 and λ_4 and their directional derivatives along the vectors of canonical basis $\mathbf{e}_p$ of the parameter space $\mathbf{h} = [h_n, h_f]$.

λ_i	Deflected shape	$\sigma_i + \omega_i$	$\dfrac{\partial\lambda_i}{\partial h_n} \equiv \mu_i(\mathbf{h}, \mathbf{e}_1)$	$\dfrac{\partial\lambda_i}{\partial h_f} \equiv \mu_i(\mathbf{h}, \mathbf{e}_2)$	$\left[\dfrac{\partial\lambda_i}{\partial h_n} \quad \dfrac{\partial\lambda_i}{\partial h_f}\right] \circ (\mathbf{e}_1 + \mathbf{e}_2) \equiv \mu_i(\mathbf{h}, \mathbf{e}_1 + \mathbf{e}_2) = \mu_i(\mathbf{h}, \mathbf{e}_1) + \mu_i(\mathbf{h}, \mathbf{e}_2)$
λ_3	antisym beam	−1.69186 +80.6252i	−0.17604 +4.19511i	−0.03307 +0.78807i	−0.20910 +4.98318i
λ_4	sym beam	−6.37071 +156.392i	−0.27743 +3.40018i	−0.11141 +1.36540i	−0.38884 +4.76550i

7. Conclusions

The attractiveness and advantages of the presented approach lie in the consistent and clear matrix notation, which significantly facilitates the software application. The disadvantage, on the other hand, is the need to simultaneously determine both the derivatives of the eigenvectors and to redetermine the derivatives of the eigenvalues due to the coupling of the two equations in Eq. (4.11).

In the algorithm for calculating only the derivatives of the eigenvalues, the derivatives of the vectors need not be determined at the same time. In other algorithms presented in the literature (see the review of methods in, e.g., (Choi *et al.*, 2004)), leading directly to the derivatives of the eigenvectors of $\mathbf{\Gamma}_j$, it is not necessary to determine μ_j at the same time – nevertheless, many of these algorithms require the determination of all modal vectors of the eigenproblem in order to express $\mathbf{\Gamma}_j$ as their linear combination. The algorithm presented here enables the selection of the eigenmodes and frequencies for which the derivatives are calculated. Moreover, the redetermination of the eigenvalue derivatives is due to the fact that the combinations of the eigenvectors $\check{\mathbf{\Psi}}^m$ and $\check{\mathbf{\Phi}}^m$ are present in Eq. (2.19), which uses the coefficients $\boldsymbol{\theta}_j^m$ being the eigenvectors for each eigenvalue derivative μ_j – Eq. (3.10).

In fact, Eq. (3.10) has a low dimension in the practice of modelling structures in FEM due to the algebraic multiples of the multiple eigenvalue $\check{\lambda}_m$. For example, for rotationally symmetric shell structures, the eigenvectors are associated with the occurrence of double eigenvalues $n_a(\check{\lambda}_m) = 2$ – this implies the need to solve Eq. (3.10) with a matrix dimension [2×2]; however, rarely, with an exceptional arrangement of the structural parameters of these shells, the dual eigenvalues coincide, resulting in an algebraic multiplicity $n_a(\check{\lambda}_m) = 4$. Such a situation may lead to multiple derivatives as eigenvectors $n_a(\mu_j) > 1$ in Eq. (3.10).

The analogous case is shown in the paper – the frame has double eigenfrequencies $n_a(\check{\lambda}_m) = 2$ related to antisymmetric and symmetric mode shapes. It is a typical case in civil engineering for flat and spatial frames. The example presented in the paper shows that the issue of derivation of derivatives for multiple eigenvalues must be conducted in a non-traditional way in general cases. This entails also the need to develop specialized software and introduce additional conditions to distinguish multiple and single eigenvalues.

References

1. Andrew, A.L., Chu, K.-W.E., & Lancaster, P. (1993). Derivatives of eigenvalues and eigenvectors of matrix functions. *SIAM Journal on Matrix Analysis and Applications*, *14*(4), 903–926. https://doi.org/10.1137/0614061
2. Balakrishnan, A.V. (1976). *Applied functional analysis. Applications of mathematics 3*. Springer.
3. Choi, K.-M., Jo, H.-K., Kim, W.-H., & Lee, I.-W. (2004). Sensitivity analysis of non-conservative eigensystems. *Journal of Sound and Vibration*, *274*(3–5), 997–1011. https://doi.org/10.1016/S0022-460X(03)00660-6
4. Day, D.M., & Walsh, T.F. (2007). *Quadratic eigenvalue problems*. Technical Report SAND2007-2072, Sandia National Laboratories. https://www.osti.gov/servlets/purl/912651
5. Garcia, S.R., & Horn, R.A. (2017). *A second course in linear algebra*. Cambridge Mathematical Textbooks. Cambridge University Press.
6. Gekeler, E.W. (2008). *Mathematical methods for mechanics. A handbook with MATLAB experiments*. Springer-Verlag, Berlin.
7. Hammarling. S., Munro, Ch.J., & Tisseur, F. (2013). An algorithm for the complete solution of quadratic eigenvalue problems. *ACM Transactions on Mathematical Software*, *39*(3), Article 18. https://doi.org/10.1145/2450153.2450156
8. Horn, R.A., & Johnson, Ch.R. (2013). *Matrix analysis* (2nd ed.). Cambridge University Press, New York.

9. Kim, D.-O., Kim, J.-T., Oh, J.-W., & Lee, I.-W. (1999a). Natural frequency and mode shape sensitivities of non-proportionally damped systems: Part I, Distinct natural frequencies. *Journal of the Computational Structural Engineering Institute of Korea*, *12*(1), 95–102.

10. Kim, D.-O., Kim, J.-T., Park, S.-K., & Lee, I.-W. (1999b). Natural frequency and mode shape sensitivities of non-proportionally damped systems: Part II, Multiple natural frequencies. *Journal of the Computational Structural Engineering Institute of Korea*, *12*(1), 103–109.

11. Krog, L.A., & Olhoff, N. (1995). Topology optimization of plate and shell structures with multiple eigenfrequencies. In N. Olhoff, & G.I.N. Rozvany (Eds.), *Structural and multidisciplinary optimization: Proceedings of The First World Congress of Structural and Multidisciplinary Optimization* (pp. 675–682). Pergamon Press.

12. Lancaster, P. (2013). Stability of linear gyroscopic systems: A review. *Linear Algebra and Its Applications*, *439*(3), 686–706. https://doi.org/10.1016/j.laa.2012.12.026

13. Lancaster, P., & Zaballa, I. (2009). Diagonalizable quadratic eigenvalue problems. *Mechanical Systems and Signal Processing*, *23*(4), 1134–1144. https://doi.org/10.1016/j.ymssp.2008.11.007

14. Lee, I.-W., Kim, D.-O., & Jung, G.-H. (1999a). Natural frequency and mode shape sensitivities of damped systems: Part I, Distinct natural frequencies. *Journal of Sound and Vibration*, *223*(3), 399–412. https://doi.org/10.1006/jsvi.1998.2129

15. Lee, I.-W., Kim, D.-O., & Jung, G.-H. (1999b). Natural frequency and mode shape sensitivities of damped systems: Part II, Multiple natural frequencies. *Journal of Sound and Vibration*, *223*(3), 413–424. https://doi.org/10.1006/jsvi.1998.2130

16. Leung, A.Y.T. (1993). *Dynamic stiffness and substructures*. Springer–Verlag.

17. Łasecka-Plura, M. (2023). A comparative study of the sensitivity analysis for systems with viscoelastic elements. *Archive of Mechanical Engineering*, *70*(1), 5–25. https://doi.org/10.24425/ame.2022.144077

18. Łasecka-Plura, M. (2024). Comprehensive sensitivity analysis of repeated eigenvalues and eigenvectors for structures with viscoelastic elements. *Acta Mechanica*, *235*(8), 5213–5238. https://doi.org/10.1007/s00707-024-03967-2

19. Martinez-Agirre, M., & Elejabarrieta, M.J. (2011). Higher order eigensensitivities-based numerical method for the harmonic analysis of viscoelastically damped structures. *International Journal for Numerical Methods in Engineering*, *88*(12), 1280–1296. https://doi.org/10.1002/nme.3222

20. Olhoff, N., Krog, L.A., & Lund, E. (1995). Optimization of multimodal structural eigenvalues. In N. Olhoff, & G.I.N. Rozvany (Eds.), *Structural and multidisciplinary optimization: Proceedings of The First World Congress of Structural and Multidisciplinary Optimization* (pp. 701–708). Pergamon Press.

21. Phuor, T., & Yoon, G.H. (2023). Eigensensitivity of damped system with distinct and repeated eigenvalues by chain rule. *International Journal for Numerical Methods in Engineering*, *124*(21), 4687–4717. https://doi.org/10.1002/nme.7331

22. Seyranian, A.P., Lund, E., & Olhoff, N. (1994). Multiple eigenvalues in structural optimization problems. *Structural Optimization*, *8*(4), 207–227. https://doi.org/10.1007/BF01742705

23. Tisseur, F., & Meerbergen, K. (2001). The quadratic eigenvalue problem. *SIAM Review*, *43*(2), 235–286. https://doi.org/10.1137/S0036144500381988

24. Wang, P., & Dai, H. (2015). Calculation of eigenpair derivatives for asymmetric damped systems with distinct and repeated eigenvalues. *International Journal for Numerical Methods in Engineering*, *103*(7), 501–515. https://doi.org/10.1002/nme.4901

25. Xu, Z., & Wu, B. (2008). Derivatives of complex eigenvectors with distinct and repeated eigenvalues. *International Journal for Numerical Methods in Engineering*, *75*(8), 945–963. https://doi.org/10.1002/nme.2280

Manuscript received November 21, 2024; accepted for publication March 19, 2025; published online September 22, 2025.

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 737–741, Warsaw 2025
https://doi.org/10.15632/jtam-pl/204808

DAMAGE LOCALIZATION IN THE MAIN STRUCTURAL ELEMENTS OF STEEL HALLS APPLYING DYNAMIC STRUCTURAL RESPONSE SIGNAL AND DISCRETE WAVELET TRANSFORM†

Anna KNITTER-PIĄTKOWSKA, Olga KAWA, Michał GUMINIAK*

Institute of Structural Analysis of Poznan University of Technology, Poznan, Poland
*corresponding author, *michal.guminiak@put.poznan.pl*

This study presents a method for detecting damage in steel lattice structures, steel hot-rolled I-section or concrete columns being a part of a whole structure based on the discrete wavelet transform (DWT). The structure's response signal may be a discrete set of displacements measured at selected points of the considered structure. The response signal defined in this way is subjected to DWT. This can have significant advantages in the field of estimation of the occurrence of a weakened part of a structure. The structural dynamic behavior is represented as a series of displacements or angles of rotation which can be implemented by given different dynamic loads, for example, short term concentrated load or seismic accelerations.

Keywords: steel lattice girders; steel welded girders; concrete bars; dynamic response signal; discrete wavelet transform.

1. Introduction – Dynamic analysis of halls

The analysis will cover steel and mixed steel-reinforced concrete elements of the structures. The analyzed constructions are real objects that are in the design and implementation phase. The signal of the structural response will be the displacements or angles of rotation of discrete points caused by a dynamic load. An example of such a load is the dynamic action generated by an earthquake. This dataset is a series of data from the 1986 Bucharest earthquake which is simulated by the computational FEM program (Axis VM). This input data set is directly implemented in the Axis VM program. It contains data in time intervals of 0.02 s, and the total duration of the function is 20.42 s. The set is dimensionless and it is the so-called load factor function. The acceleration of the support included in the calculations is determined by multiplying a given function by the (constant) acceleration value for a given direction. To obtain the values recorded at the earthquake site, an acceleration value of $1\,\mathrm{m/s^2}$ should be assumed. The FEM procedures with a bar and shell model of the structure will be utilized in the numerical analysis of the issue.

2. Foundations of the discrete wavelet transform

In the presented discrete wavelet transform (DWT) analysis, a one-dimensional wavelet transform will be used to localize defected parts of the structure. The wavelet transform has already been applied into similar problems, e.g., (Garcia-Perez *et al.*, 2013; Wang *et al.* 2013). Further,

†The content of this article was presented during the 31st Conference Vibrations in Physical and Technical Systems – VIBSYS, Poznań, Poland, October 16–18, 2024.

standard wavelet function notations are used, which were also used in previous works, e.g., (Guminiak & Knitter-Piątkowska, 2018; Kamiński *et al.*, 2025; Knitter-Piątkowska *et al.*, 2025). Let there be a given function $\psi(t)$, which is continuous and lies within the $L^2(\mathbf{R})$ space. This is the so-called mother function (wavelet function) and is required to meet the admissibility condition (Mallat, 1999). Real-valued wavelet functions are applied to the analysis. The wavelet family is derived through the translation and scaling of the mother wavelet ψ (Knitter-Piątkowska *et al.*, 2025; Guminiak & Knitter-Piątkowska, 2018):

$$\psi_{a,b} = \frac{1}{\sqrt{|a|}} \cdot \psi\left(\frac{t-b}{a}\right), \tag{2.1}$$

where t serves as a time or spatial coordinate, while a and b represent the scaling and translation parameters, respectively. The variables $(a, b \in (\mathbf{R}))$, $a \neq 0$. Finally, the DWT procedure will be applied by the substitution $a = 1/2^j$ and $b = k/2^j$ in Eq. (2.1):

$$\psi_{j,k}(t) = 2^{(j/2)} \cdot \psi\left(2^j \cdot t - k\right), \tag{2.2}$$

where k and j correspond to the scaling and translation parameters, respectively.

The DWT of the structural response data is defined through the relation

$$Wf(j,k) = 2^{j/2} \cdot \int_{-\infty}^{\infty} f(t) \cdot \psi(2^j \cdot t - k) \cdot \mathrm{d}t = \langle f(t), \psi_{j,k} \rangle \,. \tag{2.3}$$

The decomposing procedure of a signal registered in discrete points is performed while utilizing the Mallat pyramid algorithm (Mallat, 1999):

$$f_J = S_J + D_J + ... + D_n + ... + D_1, \tag{2.4}$$

where J, D_J, S_J, D_1 express in turn: the level of multiresolution analysis (MRA), the detailed and approximation components of the transformed structural response, with the most detailed one as the last term in Eq. (2.4). The Mallat pyramid algorithm is shown in Fig. 1.

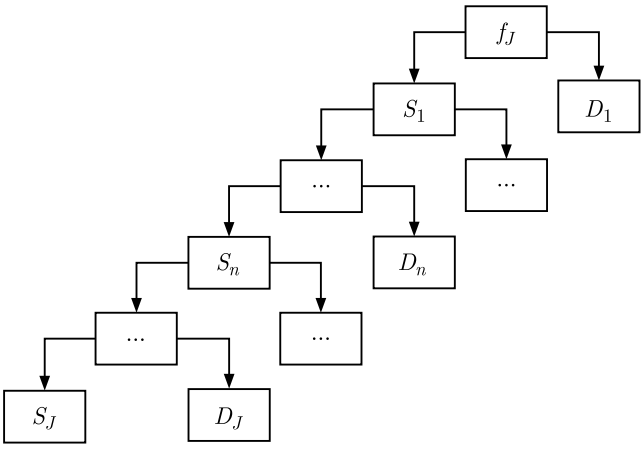

Fig. 1. Mallat pyramid algorithm.

3. Numerical examples

In this analysis, selected elements of various types of structures are considered: steel lattice and steel column, which are subjected to the dynamic load. The signal subjected to DWT is a difference between displacements or angles of rotation measured in discrete points in undamaged and defective structures.

3.1. Example 1 – Deterioration of the steel column

The steel hall with plane welded girders is considered. The arrangement of girders together with the damaged column under consideration is shown in Fig. 2. The structural properties are: width – 24 m, length – 40 m, height – 10 m, frame spacing – 5 m. The upper beam of the frame is constructed from the IPE 450 and column from the HEA 400, steel: S355 (Young's modulus 210 GPa). The introduced defect is in the form of 20 % reduction of moment of inertia in one FEM element of the column. The results of DWT calculation are shown in Fig. 3. Two-node 3D frame finite elements with six degrees of freedom per node (three mutually perpendicular translational displacements and three rotation angles in three mutually perpendicular planes) are introduced. The number of finite elements in the whole structure is 792, and in the column – 64. The structure is loaded by accelerations applied in supports along x direction.

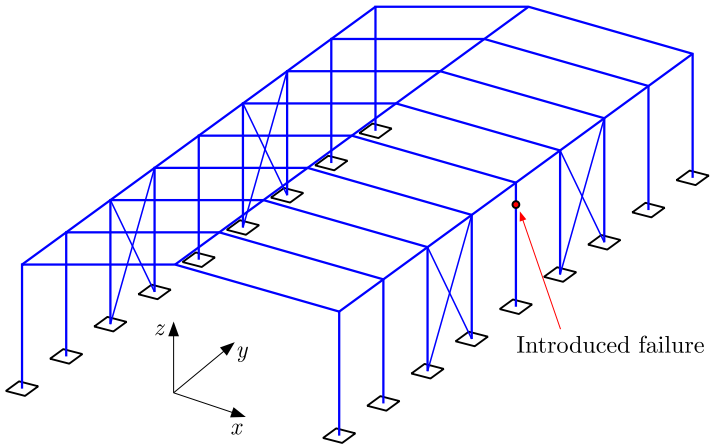

Fig. 2. Considered steel structure and introduced defective element.

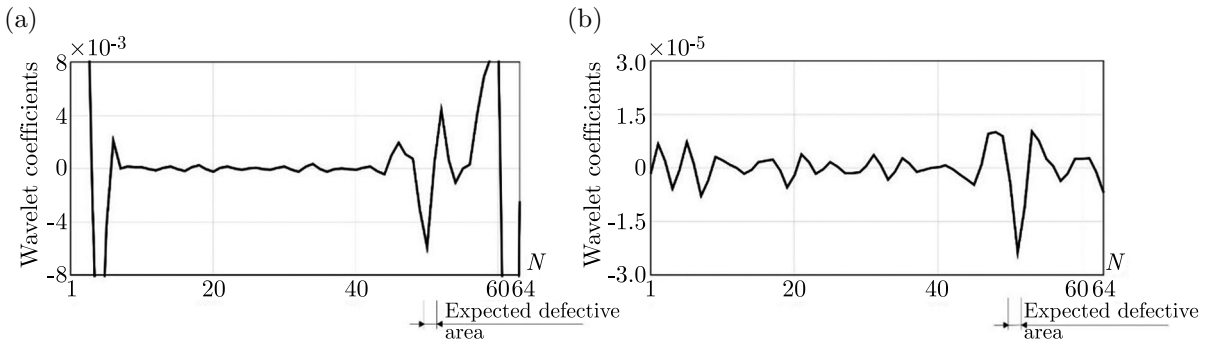

Fig. 3. Results of DWT calculations: (a) Daubechies 4, detail 1, signal – difference between displacements u_x for undamaged and defective structure; (b) Daubechies 6, detail 1, signal – difference in rotation angles φ_y (rotation in x–z plane) for undamaged and damaged structure, N – the number of measurement points.

3.2. Example 2 – Deterioration of the truss structure

The steel lattice girder resting on two reinforced concrete columns is considered. The arrangement of girders together with the damaged diagonal under consideration is shown in Fig. 4. The structural properties are: width – 24 m, length – 36 m, height – 11 m, frame spacing – 6 m. The lattice girder's upper chord is made of HEA 140, lower chord – HEA 160 and diagonals – SHS $90 \times 90 \times 5$ and steel S355 (Young's modulus 210 GPa).

The introduced defect is in the form of a broken weld between the diagonal and the lower chord, indicated in Fig. 4 by the circle. The discrete measurement points are also visible. Two-

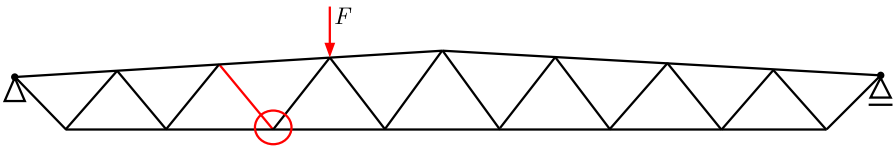

Fig. 4. Considered steel girder and introduced defective connection of diagonal bar in the node.

node frame finite elements with three degrees of freedom per node (horizontal and vertical displacements and angle of rotation) are introduced. The number of finite elements is 102, and in the lower band – 70. Loading parameters of the external harmonic force: $F = 6.67\,\text{kN}$, time of signal registration $T = 4.6\,\text{s}$. The outcomes of DTW calculation are depicted in Fig. 5.

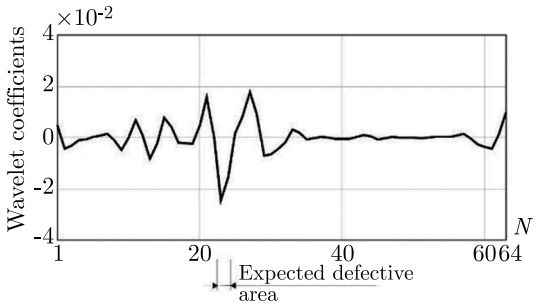

Fig. 5. Results of DWT calculations: Daubechies 6, detail 1, signal: difference in rotation angles φ for undamaged and damaged structures, N – number of measurement points.

4. Conclusions

The aim of this work was to verify the effectiveness of using discrete wavelet transform in the detection of damage in dynamically loaded structures. The calculation results presented in the previous sections allow the following conclusions to be drawn:

- DWT is highly effective for damage recognition, offering numerical efficiency and the ability to identify subtle disturbances in the response signal of a defective structure. However, in dynamic loading scenarios, referencing a signal from an undamaged structure may be necessary;
- different types of structural discrete response (i.e., translational displacements, angles of rotation) can be used for wavelet transformation in order to localize damage;
- correct damage localization may be possible regardless of the type of defect.

Acknowledgments

This research was funded by the Institute of Structural Analysis, Poznan University of Technology, internal grant number 0411/SBAD/0008.

References

1. Axis VM (in Polish), https://gammacad.pl/app/uploads/acf-uploads/axisvm_x7_podrecznik.pdf
2. Garcia-Perez, A., Amezquita-Sanchez, J.P., Dominguez-Gonzalez, A., Sedaghati, R., Osornio-Rios, R., & Romero-Troncoso, R.J. (2013). Fused empirical mode decomposition and wavelets for locating combined damage in a truss-type structure through vibration analysis. *Journal of Zhejiang University SCIENCE A. Applied Physics and Engineering*, *14*(9), 615–630. https://doi.org/10.1631/jzus.A1300030

3. Guminiak, M., & Knitter-Piątkowska, A. (2018). Selected problems of damage detection in internally supported plates using one-dimensional Discrete Wavelet Transform. *Journal of Theoretical and Applied Mechanics*, *56*(3), 631–644. https://doi.org/10.15632/jtam-pl.56.3.631

4. Kamiński, M., Guminiak, M., Knitter-Piątkowska, A., & Kawa, O. (2025). Wavelet-based stochastic finite element analysis of steel girders. *Archives of Civil Engineering*, *71*(2), 225–242. https://doi.org/10.24425/ace.2025.154118

5. Knitter-Piątkowska, A., Przychodzki, M., & Guminiak, M. (2025). Application of the Discrete Wavelet Transform to damage detection in a guy cable of guyed antenna mast. *Archives of Civil Engineering*, *71*(1), 187–201. https://doi.org/10.24425/ace.2025.153329

6. Mallat, S. (1999). *A wavelet tour of signal processing*, Academic Press: San Diego.

7. Wang, Q., Liu, H., & Wang, Q. (2013). Identification of fracture damage of the space truss structure based on the combined application of wavelet analysis and strain mode method. *Applied Mechanics and Materials*, *351–352*, 1130–1137. https://doi.org/10.4028/www.scientific.net/AMM.351-352.1130

Manuscript received December 16, 2024; accepted for publication May 7, 2025; published online August 26, 2025.

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 743–753, Warsaw 2025
https://doi.org/10.15632/jtam-pl/205104

ELASTIC BUCKLING OF AN INDIVIDUAL I-BEAM WITH CONSIDERATION OF THE SHEAR EFFECT

Krzysztof MAGNUCKI

Łukasiewicz Research Network, Poznan Institute of Technology, Poznan, Poland
krzysztof.magnucki@pit.lukasiewicz.gov.pl

The subject of the paper is a homogeneous I-beam with an individual shape web. The beam, simply supported at one end and at the other simply supported with elastic limitation of rotation, is subjected to axial compression by a force F. The analytical model of the beam is developed with consideration of the shear effect. The deformation of the beam's plane cross-section after buckling is determined analytically, taking into account the classical expression for shear stresses in a beam (known as Zhuravsky or Jourawski shear stress). Longitudinal displacements, strains, and stresses are then formulated. Based on the principle of stationary total potential energy, a system of two equilibrium differential equations is derived. These equations are solved analytically, taking into account the beam support conditions, and the critical force FCR is determined. Detailed calculations are realized for sample beams.

Keywords: analytical modeling; I-beam; elastic buckling; shear effect.

1. Introduction

Initiated by Leonhard Euler in the 18th century, the study of buckling of a compressed beam has been intensively developed over the following centuries and is still being improved today to address problems related to the stability of beams, plates, and shells, which are closely related to the design of structures. Rykaluk (2012) described in detail the criteria of elastic stability, buckling issues of axially and eccentrically compressed classical and thin-walled beams with open cross-sections, buckling of rings and arches, stability of flat frames, as well as stability problems of rectangular plates and shells. Magnucki and Milecki (2015) examined the elastic buckling problem of the symmetrical triangular frame under tensile in-plane load. They studied in detail both analytically and numerically (using FEM) the in-plane buckling state and the lateral buckling state of this sample frame. Simão (2017) presented a stability analysis of shear-sensitive columns with a linear formulation according to the Timoshenko beam theory along with a nonlinear shear formula. Eslami (2018) first characterized the stability concept and then presented in detail buckling and post-buckling problems of beams and plates, as well as buckling problems of cylindrical, spherical and conical shells. Yang *et al.* (2019) presented the results of multiple numerical and experimental studies of the global buckling problem of bi-symmetrical steel beams under three-point bending. Szymczak and Kujawa (2019) investigated analytically and

Ministry of Science and Higher Education
Republic of Poland

This publication has been funded by the Polish Ministry of Science and Higher Education under the Excellent Science II programme "Support for scientific conferences".
The content of this article was presented during the 61st Symposium "Modelowanie w mechanice" (Modelling in Mechanics), Szczyrk, Poland, March 2–5, 2025.

numerically (FEM) the flexural buckling of axially compressed, simply supported, and clamped I-columns made of aluminum alloy and indicated the influence of material non-linearity on the critical loads. Genovese and Elishakoff (2019) pointed out the importance of the principle of virtual work in the formulation of planar static rod theories with consideration of large deformations and the transverse shear effect. Filho *et al.* (2022) investigated, both experimentally and numerically, the buckling problem of welded I-section columns undergoing flexural or torsional buckling failure. Yang *et al.* (2023) focused on the buckling instability problem of I-beams, developed a numerical model of the beam using Lagrange polynomials to describe the three-dimensional displacement field, and numerically investigated the global buckling, local buckling, and global–local coupled buckling of these beams. Jing *et al.* (2024) studied the effect of limiting the beam center deflection with an elastic support on the critical axial force and dynamic characteristics. They determined two stable states of beam buckling depending on the stiffness of the elastic center support.

Magnucki and Sowiński (2024) applied an individual nonlinear shear deformation theory to the analytical modeling of a clamped sandwich beam with a functionally graded core, and then analytically and numerically (using FEM) studied the bending of this beam under a uniformly distributed load. Magnucki (2024a) developed two analytical models of a five-layered composite beam. The first is formulated on the basis of the classical zig-zag theory, while the second is developed using the nonlinear shear deformation theory. He then analytically studied the bending of these sample beams for both models. Magnucki (2024b) analytically described the cross-section of a standard wide-flange H-beam as a three-layer structure and analytically studied the fundamental natural frequency of these sample beams with consideration of the shear effect. Couto *et al.* (2025) provided a thorough review of the research that led to the proposal of European fire design rules for steel thin-walled I-beams. They focused on the problem of interaction between local and global buckling in these members.

This work continues studies of the beam buckling problems presented in the above sample papers. The main goal of the work is to analytically study the beam buckling with consideration of the shear effect.

2. Analytical study of the individual I-beam with consideration of the shear effect

The subject of the study is a homogeneous individual I-beam of length L, width b, and total depth h under axial compression by a force F. One end of the beam is simply supported, while the other end is elastically limited in rotation by means of a rigid part connected to two springs with stiffness k_s (Fig. 1). The beam is protected against out of the x- and y-plane buckling.

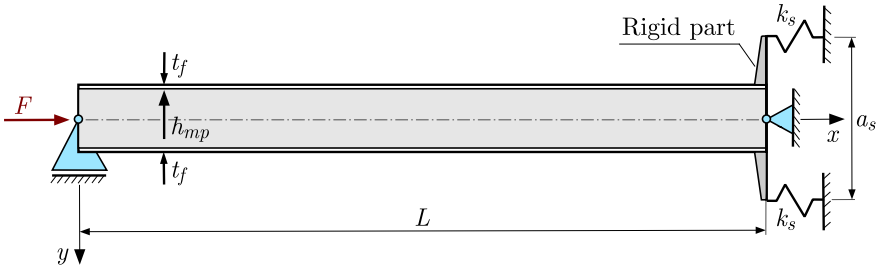

Fig. 1. Schematic diagram of the beam with two different supports at its ends.

The cross-section of this beam, with a functionally graded middle part-web thickness, is shown in Fig. 2. This individual beam's planar cross-section provides input for analytically determining the nonlinear function of its deformation.

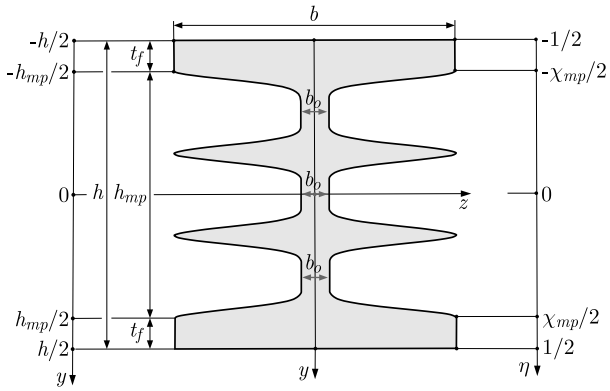

Fig. 2. Schematic cross-section of the individual beam.

The widths of the successive parts of this cross-section are as follows:

– the upper flange ($-1/2 \leq \eta \leq -\chi_{mp}/2$):

$$\overline{w}_{uf}(\eta) = \frac{w_{uf}(\eta)}{b} = 1, \tag{2.1}$$

– the middle part – web ($-\chi_{mp}/2 \leq \eta \leq \chi_{mp}/2$):

$$\overline{w}_{mp}(\eta) = \frac{w_{mp}(\eta)}{b} = \beta_0 + (1-\beta_0)\sin^n\left(3\pi\frac{\eta}{\chi_{mp}}\right), \tag{2.2}$$

– the lower flange ($\chi_{mp}/2 \leq \eta \leq 1/2$):

$$\overline{w}_{lf}(\eta) = \frac{w_{lf}(\eta)}{b} = 1, \tag{2.3}$$

where $\eta = y/h$ – dimensionless coordinate, $\chi_{mp} = h_{mp}/h$, $\beta_0 = b_0/b$ – dimensionless sizes, and n – even number.

Taking into account the papers by Magnucki (2024a) and Magnucki (2024b), the dimensionless deformation functions for the successive parts of this cross-section are analytically determined as follows:

– the upper flange ($-1/2 \leq \eta \leq -\chi_{mp}/2$):

The dimensionless first moment (Fig. 3) is given by

$$\overline{S}_z^{(uf)}(\eta) = \frac{S_z^{(uf)}(\eta)}{bh^2} = -\int_{-1/2}^{-\eta} \eta\,\mathrm{d}\eta = \frac{1}{8}(1-4\eta^2). \tag{2.4}$$

Therefore, the derivative of the dimensionless deformation function is

$$\frac{\mathrm{d}f_d^{(uf)}}{\mathrm{d}\eta} = \frac{\overline{S}_z^{(uf)}(\eta)}{\overline{w}_{uf}(\eta)} = \frac{1}{8}(1-4\eta^2). \tag{2.5}$$

Consequently, the dimensionless deformation function is

$$f_d^{(uf)}(\eta) = -C_f + \frac{1}{8}\left(1-\frac{4}{3}\eta^2\right)\eta, \tag{2.6}$$

where the integration constant:

$$C_f = -\frac{1}{16}\left(1-\frac{1}{3}\chi_{mp}^2\right)\chi_{mp} + \int_0^{\chi_{mp}/2} \frac{\overline{S}_z^{(mp)}(\eta)}{\overline{w}_{mp}(\eta)}\,\mathrm{d}\eta; \tag{2.7}$$

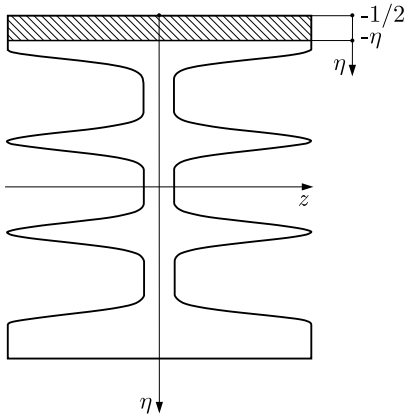

Fig. 3. Hatched area of the upper flange selected part.

– the middle part–web ($-\chi_{mp}/2 \leq \eta \leq \chi_{mp}/2$):
The dimensionless first moment (Fig. 4) is given by

$$\overline{S}_z^{(mp)}(\eta) = \frac{1}{8}\left(1-\chi_{mp}^2\right) - \int\limits_{-\chi_{mp}/2}^{\eta} \eta \overline{w}_{mp}(\eta)\,\mathrm{d}\eta. \tag{2.8}$$

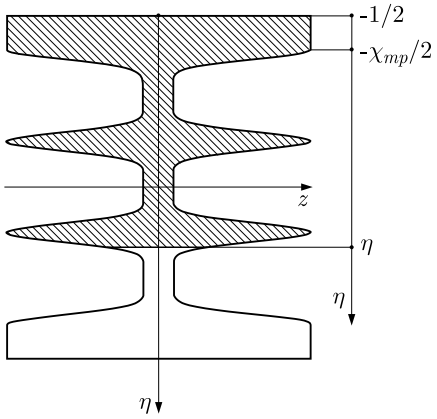

Fig. 4. Hatched area of the selected part of the functionally graded middle part-web.

The derivative of the dimensionless deformation function is

$$\frac{\mathrm{d}f_d^{(mp)}}{\mathrm{d}\eta} = \frac{\overline{S}_z^{(mp)}(\eta)}{\overline{w}_{mp}(\eta)} \tag{2.9}$$

and the dimensionless deformation function is

$$f_d^{(mp)}(\eta) = \int \frac{\overline{S}_z^{(mp)}(\eta)}{\overline{w}_{mp}(\eta)}\,\mathrm{d}\eta; \tag{2.10}$$

– the lower flange ($\chi_{mp}/2 \leq \eta \leq 1/2$):

$$\overline{S}_z^{(lf)}(\eta) = \frac{1}{8}(1-4\eta^2), \tag{2.11}$$

$$\frac{\mathrm{d}f_d^{(lf)}}{\mathrm{d}\eta} = \frac{1}{8}(1-4\eta^2), \tag{2.12}$$

$$f_d^{(lf)}(\eta) = C_f + \frac{1}{8}\left(1-\frac{4}{3}\eta^2\right)\eta. \tag{2.13}$$

The schematic cross-section of the beam, and graphs of dimensionless deformation functions of the planar cross-section (2.6), (2.10), (2.13), and its derivatives (2.5), (2.9), (2.12) for the example beam ($\chi_{mp} = 4/5$, $\beta_0 = 1/10$, $n = 10$) are shown in Fig. 5.

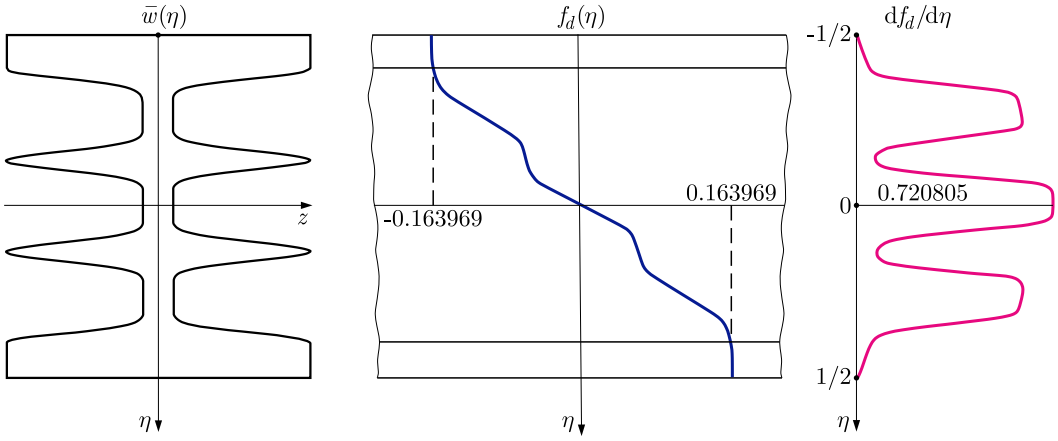

Fig. 5. Schematic cross-section of the beam and graphs of functions $f_d(\eta)$, $\mathrm{d}f_d/\mathrm{d}\eta$.

The deformation of a planar cross-section of this beam, in accordance with the nonlinear shear deformation theory, is shown in Fig. 6.

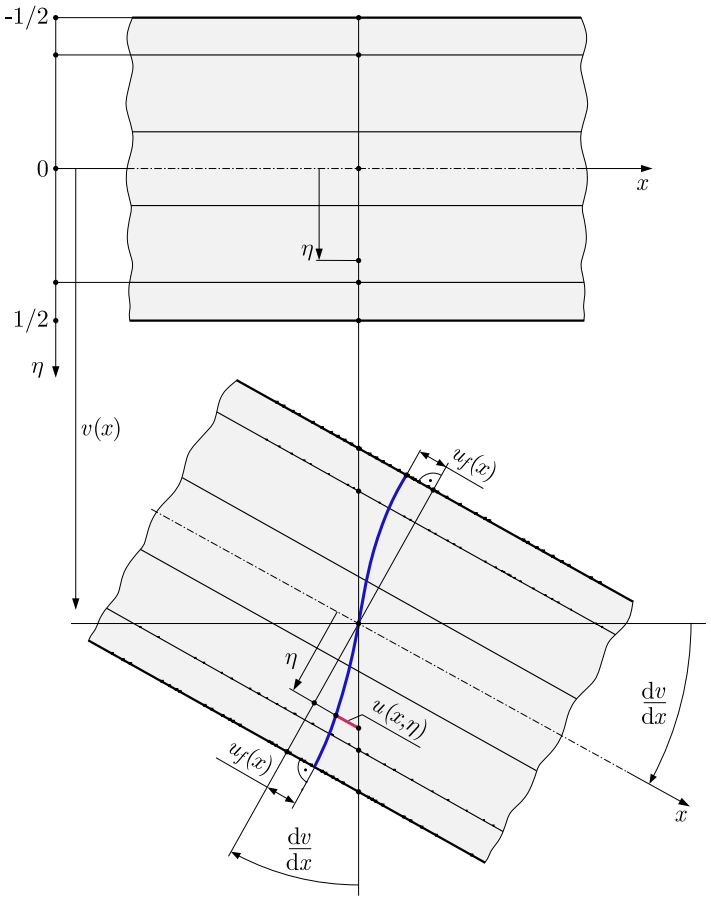

Fig. 6. Schematic diagram of planar cross-section deformation of the beam.

Based on this diagram (Fig. 6), the longitudinal displacements, and consequently strains and stresses, in the successive parts of this cross-section are as follows:

– upper flange ($-1/2 \leq \eta \leq -\chi_{mp}/2$):

$$u^{(uf)}(x,\eta) = -h\left[\eta\frac{\mathrm{d}v}{\mathrm{d}x} - f_d^{(uf)}(\eta)\psi_f(x)\right], \tag{2.14}$$

$$\varepsilon_x^{(uf)}(x,\eta) = -h\left[\eta\frac{\mathrm{d}^2v}{\mathrm{d}x^2} - f_d^{(uf)}(\eta)\frac{\mathrm{d}\psi_f}{\mathrm{d}x}\right], \tag{2.15}$$

$$\gamma_{xy}^{(uf)}(x,\eta) = \frac{\mathrm{d}f_d^{(uf)}}{\mathrm{d}\eta}\psi_f(x), \tag{2.16}$$

$$\sigma_x^{(uf)}(x,\eta) = E\,\varepsilon_x^{(uf)}(x,\eta), \tag{2.17}$$

$$\tau_{xy}^{(uf)}(x,\eta) = \frac{E}{2(1+\nu)}\gamma_{xy}^{(uf)}(x,\eta), \tag{2.18}$$

– the middle part – web ($-\chi_{mp}/2 \leq \eta \leq \chi_{mp}/2$):

$$u^{(mp)}(x,\eta) = -h\left[\eta\frac{\mathrm{d}v}{\mathrm{d}x} - f_d^{(mp)}(\eta)\psi_f(x)\right], \tag{2.19}$$

$$\varepsilon_x^{(mp)}(x,\eta) = -h\left[\eta\frac{\mathrm{d}^2v}{\mathrm{d}x^2} - f_d^{(mp)}(\eta)\frac{\mathrm{d}\psi_f}{\mathrm{d}x}\right], \tag{2.20}$$

$$\gamma_{xy}^{(mp)}(x,\eta) = \frac{\mathrm{d}f_d^{(mp)}}{\mathrm{d}\eta}\psi_f(x), \tag{2.21}$$

$$\sigma_x^{(mp)}(x,\eta) = E\,\varepsilon_x^{(mp)}(x,\eta), \tag{2.22}$$

$$\tau_{xy}^{(mp)}(x,\eta) = \frac{E}{2(1+\nu)}\gamma_{xy}^{(mp)}(x,\eta), \tag{2.23}$$

– the lower flange ($\chi_{mp}/2 \leq \eta \leq 1/2$):

$$u^{(lf)}(x,\eta) = -h\left[\eta\frac{\mathrm{d}v}{\mathrm{d}x} - f_d^{(lf)}(\eta)\psi_f(x)\right], \tag{2.24}$$

$$\varepsilon_x^{(lf)}(x,\eta) = -h\left[\eta\frac{\mathrm{d}^2v}{\mathrm{d}x^2} - f_d^{(lf)}(\eta)\frac{\mathrm{d}\psi_f}{\mathrm{d}x}\right], \tag{2.25}$$

$$\gamma_{xy}^{(lf)}(x,\eta) = \frac{\mathrm{d}f_d^{(lf)}}{\mathrm{d}\eta}\psi_f(x), \tag{2.26}$$

$$\sigma_x^{(lf)}(x,\eta) = E\,\varepsilon_x^{(lf)}(x,\eta), \tag{2.27}$$

$$\tau_{xy}^{(lf)}(x,\eta) = \frac{E}{2(1+\nu)}\gamma_{xy}^{(lf)}(x,\eta), \tag{2.28}$$

where E – Young's modulus, ν – Poisson's ratio.

Based on the principle of stationary total potential energy, after simple transformation, the system of two differential equilibrium equations for this beam is obtained in the following form:

$$\overline{J}_z \frac{\mathrm{d}^2 v}{\mathrm{d}x^2} - C_{v\psi} \frac{\mathrm{d}\psi_f}{\mathrm{d}x} = -\frac{M_b(x)}{Ebh^3}, \tag{2.29}$$

$$C_{v\psi} \frac{\mathrm{d}^3 v}{\mathrm{d}x^3} - C_{\psi\psi} \frac{\mathrm{d}^2 \psi_f}{\mathrm{d}x^2} + C_{\psi} \frac{\psi_f(x)}{h^2} = 0, \tag{2.30}$$

where dimensionless coefficients:

$$\overline{J}_z = \frac{1}{12}\left[1 - (1-\beta_0)\,\chi_{mp}^3\right] + (1-\beta_0) \int\limits_{-\chi_{mp}/2}^{\chi_{mp}/2} \eta^2 \sin^n\left(3\pi \frac{\eta}{\chi_{mp}}\right) \mathrm{d}\eta,$$

$$C_{v\psi} = 2\left\{ \int\limits_{\chi_{mp}/2}^{1/2} \eta\, f_d^{(lf)}(\eta)\,\mathrm{d}\eta + \int\limits_{0}^{\chi_{mp}/2} \eta\, f_d^{(mp)}(\eta)\; \overline{w}_{mp}(\eta)\,\mathrm{d}\eta \right\},$$

$$C_{\psi\psi} = 2\left\{ \int\limits_{\chi_{mp}/2}^{1/2} \left[f_d^{(lf)}(\eta)\right]^2 \mathrm{d}\eta + \int\limits_{0}^{\chi_{mp}/2} \left[f_d^{(mp)}(\eta)\right]^2 \overline{w}_{mp}(\eta)\,\mathrm{d}\eta \right\},$$

$$C_{\psi} = \frac{1}{1+\nu}\left\{ \int\limits_{\chi_{mp}/2}^{1/2} \left(\frac{\mathrm{d} f_d^{(lf)}}{\mathrm{d}\eta}\right)^2 \mathrm{d}\eta + \int\limits_{0}^{\chi_{mp}/2} \left(\frac{\mathrm{d} f_d^{(mp)}}{\mathrm{d}\eta}\right)^2 \overline{w}_{mp}(\eta)\,\mathrm{d}\eta \right\}.$$

The buckled shape of this beam is shown in Fig. 7.

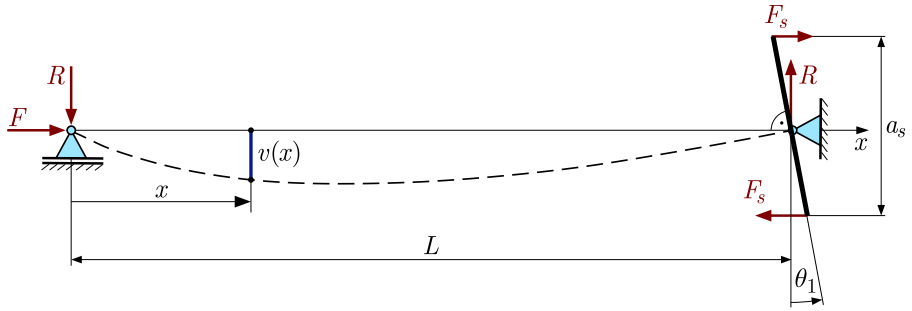

Fig. 7. Schematic diagram of the buckled shape of this beam.

The bending moment according to this diagram takes the following form:

$$M_b(x) = \left[-\overline{R}x + \overline{F}v(x)\right] Ebh, \tag{2.31}$$

where $\overline{R} = R/Ebh$, $\overline{F} = F/Ebh$ – dimensionless reaction and axial force.

Thus, the two differential equations of equilibrium (2.29) and (2.30) in the dimensionless coordinate, with consideration of expression (2.31), are as follows:

$$\overline{J}_z \frac{\mathrm{d}^2 v}{\mathrm{d}\xi^2} - C_{v\psi} L \frac{\mathrm{d}\psi_f}{\mathrm{d}\xi} + \lambda^2 \overline{F} v(\xi) = \xi \lambda^2 L \overline{R}, \tag{2.32}$$

$$C_{v\psi} \frac{\mathrm{d}^3 v}{\mathrm{d}\xi^3} - C_{\psi\psi} L \frac{\mathrm{d}^2 \psi_f}{\mathrm{d}\xi^2} + C_{\psi} \lambda^2 L \psi_f(\xi) = 0. \tag{2.33}$$

These two equations, after simple transformations, are reduced to one in the following form:

$$\frac{\mathrm{d}^4 v}{\mathrm{d}\xi^4} - k_2 \frac{\mathrm{d}^2 v}{\mathrm{d}\xi^2} - k_0 v(\xi) = -\xi\, \alpha^2 \frac{\lambda^4}{\overline{J}_z} \overline{R} L, \tag{2.34}$$

where $k_0 = \alpha^2 \frac{\lambda^4}{\overline{J}_z} \overline{F}$, $k_2 = \alpha^2 \lambda^2 \left(1 - \frac{C_{\psi\psi}}{\overline{J}_z C_{\psi}} \overline{F}\right)$, $\alpha = \sqrt{\frac{\overline{J}_z C_{\psi}}{\overline{J}_z C_{\psi\psi} - C_{v\psi}^2}}$ – dimensionless coefficients.

The solution of Eq. (2.34), with consideration of the boundary conditions, $v(0) = v(1) = 0$, is in the following form:

$$v(\xi) = \left[\xi - \frac{\sin(q\xi)}{\sin(q)}\right] \frac{\overline{R}}{\overline{F}} L, \tag{2.35}$$

where $q = \frac{1}{\sqrt{2}} \sqrt{-k_2 + \sqrt{k_2^2 + 4k_0}}$ – the dimensionless coefficient.

This function (2.35) describes the buckling line of the beam. Thus, the derivative of this function is given by

$$\frac{\mathrm{d}v}{\mathrm{d}\xi} = \left[1 - q \frac{\cos(q\xi)}{\sin(q)}\right] \frac{\overline{R}}{\overline{F}} L. \tag{2.36}$$

Consequently, the angle of rotation of the simply supported end of this beam ($\xi = 1$) with rotation limitation takes the following form:

$$\theta_1 = \left.\frac{\mathrm{d}v}{L\,\mathrm{d}\xi}\right|_1 = \left[1 - \frac{q}{\tan(q)}\right] \frac{\overline{R}}{\overline{F}}. \tag{2.37}$$

Based on Fig. 1 and Fig. 7, two expressions are formulated: $F_s a_s = RL$ and $F_s = \frac{1}{2}\theta_1 k_s a_s$, from which the angle rotation

$$\theta_1 = k_{\theta 1} \overline{R}, \tag{2.38}$$

where $k_{\theta 1} = \frac{2EbhL}{k_s a_s^2}$ – dimensionless coefficient, k_s – spring stiffness [N/mm], a_s – size [mm].

Equating both expressions (2.37) and (2.38), one obtains the algebraic equation

$$1 - \frac{q}{\tan(q)} + k_{\theta 1} \overline{F} = 0. \tag{2.39}$$

Based on this equation, the dimensionless critical force $\overline{F}_{CR}$ of this beam is determined.

Analyzing this equation, it is easy to notice two classic cases of support:

1) a simply supported beam for $k_s = 0$, $k_{\theta 1} \to \infty$, then $\tan(q) = 0$, from which $q = \pi$,
2) one clamped end for $k_s \to \infty$, $k_{\theta 1} = 0$, then $\tan(q) = q$, from which $q \cong 4.4934$.

Example calculations are carried out for the following data: $\chi_{mp} = 4/5$, $\beta_0 = 1/10$, $n = 10$, $\nu = 0.3$, $\lambda = 40$, and the dimensionless coefficient ($0 \le k_{\theta 1} < \infty$). The results of analytical calculations of the selected values of the dimensionless critical force $\overline{F}_{CR}$ are specified in Table 1.

Table 1. Values of the dimensionless critical force $\overline{F}_{CR}$ for the compressed beam.

$k_{\theta 1}$	0	2000	4000	7000	15 000	30 000	∞
$10^4 \overline{F}_{CR}$	6.93366	6.14493	5.60592	5.08517	4.42156	3.98617	3.41070

Moreover, these critical force values are graphically presented in Fig. 8.

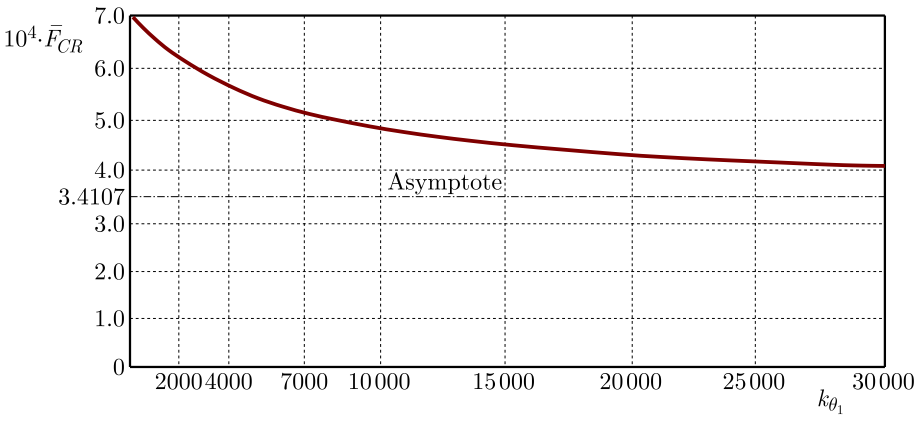

Fig. 8. Graph of the dimensionless critical force.

3. Analytical study of the individual I-beam without the shear effect

The two differential equilibrium Eqs. (2.32) and (2.33), in this case, are reduced to one in the following form:

$$\frac{\mathrm{d}^2 v}{\mathrm{d}\xi^2} + \frac{\overline{F}}{\overline{J}_z}\lambda^2 v(\xi) = \xi \frac{\overline{R}}{\overline{J}_z}\lambda^2 L. \tag{3.1}$$

The solution of Eq. (3.1), with consideration of the boundary conditions, $v(0) = v(1) = 0$, is in the following form:

$$v(\xi) = \left[\xi - \frac{\sin(k_F \xi)}{\sin(k_F)}\right] \frac{\overline{R}}{\overline{F}} L, \tag{3.2}$$

where $k_F = \sqrt{\overline{F}/\overline{J}_z}\lambda$ – dimensionless coefficient.

The derivative of this function is given by:

$$\frac{\mathrm{d}v}{\mathrm{d}\xi} = \left[1 - k_F \frac{\cos(k_F \xi)}{\sin(k_F)}\right] \frac{\overline{R}}{\overline{F}} L. \tag{3.3}$$

Consequently, the angle of rotation of the simply supported end of this beam ($\xi = 1$), with rotation limitation, takes the following form:

$$\theta_1 = \left.\frac{\mathrm{d}v}{L\,\mathrm{d}\xi}\right|_1 = \left[1 - \frac{k_F}{\tan(k_F)}\right] \frac{\overline{R}}{\overline{F}}. \tag{3.4}$$

This angle of rotation is consistent with the angle (2.38); therefore, by equating these two angles, one obtains the algebraic equation:

$$1 - \frac{k_F}{\tan(k_F)} + k_{\theta 1}\overline{F} = 0. \tag{3.5}$$

Based on this equation, the dimensionless critical force $\overline{F}_{CR}^{(0)}$ of this beam without the shear effect is determined. Similarly to the beam that considers the shear effect, when analyzing Eq. (3.5) for the two classic support cases, one obtains:

1) simply supported beam for $k_s = 0$, $k_{\theta 1} \to \infty$, then $\tan(k_F) = 0$, from which $k_F = \pi$, and so $\overline{F}_{CR}^{(0)} = \frac{\pi^2 \overline{J}_z}{\lambda^2}$,

2) one clamped end for $k_s \to \infty$, $k_{\theta 1} = 0$, then $\tan(k_F) = k_F$, from which $k_F \cong 4.4934$, and so $\overline{F}_{CR}^{(0)} = \frac{\pi^2 \overline{J}_z}{(0.699\lambda)^2}$.

Example calculations are carried out for the same data as for the beam with the shear effect. The results of the analytical calculations of the selected values of the dimensionless critical force $\overline{F}_{CR}^{(0)}$ are specified in Table 2.

Table 2. Values of the dimensionless critical force $\overline{F}_{CR}^{(0)}$ for the compressed beam.

$k_{\theta 1}$	0	2000	4000	7000	15 000	30 000	∞
$10^4 \overline{F}_{CR}^{(0)}$	7.01980	6.20522	5.65306	5.12268	4.45039	4.01093	3.43141

Taking into account the dimensionless critical force $\overline{F}_{CR}$ for the beam with the shear effect and the dimensionless critical force $\overline{F}_{CR}^{(0)}$ for the beam without the shear effect, the dimensionless shear effect coefficient of this beam is formulated as follows:

$$C_{SE} = 1 - \frac{\overline{F}_{CR}}{\overline{F}_{CR}^{(0)}}. \tag{3.6}$$

The example values of this coefficient are specified in Table 3.

Table 3. Values of the dimensionless shear effect coefficient C_{SE} for the beam.

$k_{\theta 1}$	0	2000	4000	7000	15 000	30 000	∞
$10^3 C_{SE}$	12.2710	9.7160	8.3388	7.3223	6.4781	6.1731	6.0354

Moreover, these dimensionless coefficient values of the shear effect are graphically presented in Fig. 9.

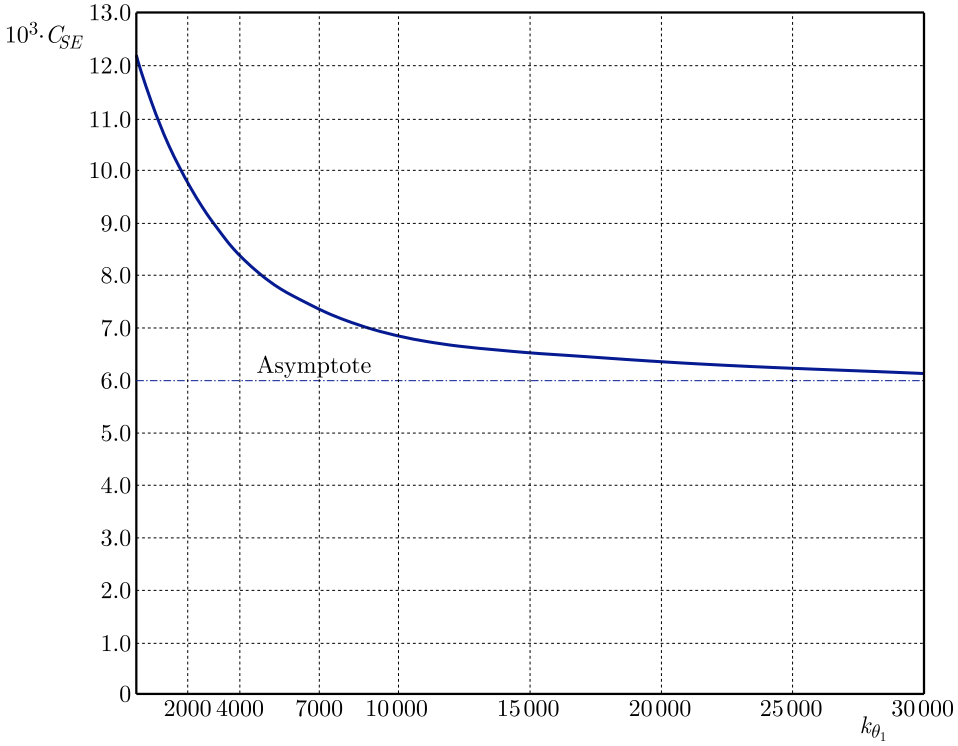

Fig. 9. Graph of the dimensionless coefficient of the shear effect.

The value of this dimensionless coefficient C_{SE} is small, so the influence of the shear effect on the critical force is insignificant.

4. Conclusions

The presented analytical studies of the elastic buckling problem for the individual I-beam, simply supported at the first end and with limited rotation at the second end, allow the following conclusions to be formulated:

- the applied procedure for analytically determining the deformation function of a planar beam cross-section is easy and effective (expressions (2.4)–(2.13)), (Figs. 3–5);
- the two different supports adopted at the beam ends make it possible to carry out tests on beams ranging from simply supported to clamped at one end (Table 1 and Fig. 8);
- the influence of the shear effect on the critical force values for the tested beam family is insignificant (Table 3 and Fig. 9).

References

1. Couto, C., Real, P.V., & Lopes, N. (2025). Local-global buckling interaction in steel I-beams—A European design proposal for the case of fire, *Thin–Walled Structures*, *206*, Part A, Article 112664. https://doi.org/10.1016/j.tws.2024.112664
2. Eslami, M.R. (2018). *Buckling and postbuckling of beams, plates, and shells*. Springer.
3. Filho, J.O.F., Tankova, T., Carvalho, H., Martins, C., & da Silva, L.S. (2022). Experimental and numerical flexural buckling resistance of high strength steel columns and beam-columns. *Engineering Structures*, *265*, Article 114414. https://doi.org/10.1016/j.engstruct.2022.114414
4. Genovese, D., & Elishakoff, I. (2019). Shear deformable rod theories and fundamental principles of mechanics. *Archive of Applied Mechanics*, *89*, 1995–2003. https://doi.org/10.1007/s00419-019-01556-7
5. Jing, J., Shao, Z.-H., Mao, X.-Y., Ding, H., & Chen, L.-Q. (2024). Forced resonance of a buckled beam flexibly restrained at the inner point. *Engineering Structures*, *303*, Article 117444. https://doi.org/10.1016/j.engstruct.2024.117444
6. Magnucki, K. (2024a). Bending of a five-layered composite beam with consideration of two analytical models. *Archive of Mechanical Engineering*, *71*(1), 27–46. https://doi.org/10.24425/ame.2024.149188
7. Magnucki, K. (2024b). Free flexural vibrations of standard wide-flange H-beams with consideration of the shear effect. *Rail Vehicles/Pojazdy Szynowe*, *1–2*, 46–50. https://doi.org/10.53502/RAIL-189244
8. Magnucki, K., & Milecki, S. (2015). Elastic buckling of a triangular frame subject to in-plane tension. *Journal of Theoretical and Applied Mechanics*, *53*(3), 581–591. http://doi.org/10.15632/jtam-pl.53.3.581
9. Magnucki, K., & Sowiński, K. (2024). Bending of a sandwich beam with an individual functionally graded core. *Journal of Theoretical and Applied Mechanics*, *62*(1), 3–17. https://doi.org/10.15632/jtam-pl/174698
10. Rykaluk, K. (2012). *Stability problems of metal constructions* (in Polish), Dolnośląskie Wydawnictwo Edukacyjne, Wrocław.
11. Simão, P.D. (2017). Influence of shear deformations on the buckling of columns using the Generalized Beam Theory and energy principles. *European Journal of Mechanics – A/Solids*, *61*, 216–234. https://doi.org/10.1016/j.euromechsol.2016.09.015
12. Szymczak, C., & Kujawa, M. (2019). Flexural buckling and post-buckling of columns made of aluminium alloy. *European Journal of Mechanics – A/Solids*, *73*, 420–429. https://doi.org/10.1016/j.euromechsol.2018.10.006
13. Yang, B., Zhang, Y., Xiong, G., Elchalakani, M., & Kang, S.-B. (2019). Global buckling investigation on laterally-unrestrained Q460GJ steel beams under three-point bending. *Engineering Structures*, *181*, 271–280. https://doi.org/10.1016/j.engstruct.2018.12.028
14. Yang, Y., Hui, Y., Li, P., Yang, J., Huang, Q., Giunta, G., Belouettar, S., & Hu, H. (2023). Global/local buckling analysis of thin-walled I-section beams via hierarchical one-dimensional finite elements. *Engineering Structures*, *280*, Article 115705. https://doi.org/10.1016/j.engstruct.2023.115705

Manuscript received March 13, 2025; accepted for publication May 13, 2025; published online September 4, 2025.

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 755–767, Warsaw 2025
https://doi.org/10.15632/jtam-pl/207298

IMPACT ACCELERATION ACQUISITION WITH A HIGH FREQUENCY DATA LOGGING SYSTEM BASED ON SPI COMMUNICATION PROTOCOL

Kamil KURPANIK, Jonasz HARTWICH, Sławomir KCIUK, Sławomir DUDA

Faculty of Mechanical Engineering Technology, Silesian University of Technology, Gliwice, Poland
*corresponding author, *kamil.kurpanik@polsl.pl*

As part of the research conducted, the correctness of the serial peripheral interface (SPI) communication was tested using the Raspberry Pi 4B. The purpose of the experiment was to confirm the possibility of stable data exchange between the control unit and external peripheral circuits, with particular emphasis on analog-to-digital converters and micro-electro-mechanical systems (MEMS) sensors. The tests showed correct and stable operation of the SPI bus for both unidirectional and bidirectional transmission, and confirmed the usefulness of SPI in measurement systems requiring precise and fast data transfer. The experiments provide a basis for further work with real-time sensors and data acquisition systems under dynamic conditions.

Keywords: data logging system; MEMS; SPI; impact acceleration; high frequency.

1. Introduction

The recording of shock wave parameters is the basis of a significant amount of research aimed at improving the safety of military vehicle crews (Baranowski *et al.*, 2020; Pyka *et al.*, 2025). A significant parameter measured in such experiments is the acceleration, which can reach a value of several hundred g in less than a few milliseconds (Kciuk *et al.*, 2022). It is particularly important to register this parameter since the forces resulting from these accelerations through the impact of the vehicle's structure are transmitted directly to the crew, which can lead to death or serious injuries such as spinal cord fracture (Elsayed & Atkins, 2008). Measurements of rapidly changing acceleration values are also used in civilian industry, an example of this type of research is car crash tests (Dima & Covaciu, 2017; Olvey *et al.*, 2004). The recording of shock wave parameters also provides a reference for numerical simulations (Arkusz *et al.*, 2019; Erdik *et al.*, 2016; Wrazidło *et al.*, 2025).

The most commonly used sensors for acceleration measurements are devices based on micro-electro-mechanical systems (MEMS) technology. These are defined as minimized mechanical and electromechanical components (with dimensions of a few hundred micrometers at most) (Covaciu & Dima, 2017). An accelerometer is essentially a capacitive or piezoresistive device consisting of a suspended pendulum proof mass/plate assembly. In the case of capacitive accelerometers, the displacement of the test mass induces a change in capacitance of the capacitor, allowing the determination of acceleration (Sethuramalingam & Vimalajuliet, 2010). The primary parameters

Ministry of Science and Higher Education
Republic of Poland

This publication has been funded by the Polish Ministry of Science and Higher Education under the Excellent Science II programme "Support for scientific conferences".
The content of this article was presented during the 61st Symposium "Modelowanie w mechanice" (Modelling in Mechanics), Szczyrk, Poland, March 2–5, 2025.

that define the applicability of a given accelerometer within a specific application are the output range and measurement resolution. Nonetheless, in the context of research investigating the impact of shock waves on human health, two crucial factors must be given due consideration. Firstly, the frequency bandwidth carried by the sensor and, secondly, the sampling frequency of the entire measurement system. In order to record a peak of acceleration values lasting only a few milliseconds during the measurement process, the sampling frequency must be sufficiently high. In the process of designing the measurement system, particular attention must be paid to the digitization of the voltage signal from the MEMS-type sensor, as well as to the communication protocols.

The signal from the MEMS-type accelerometer after sampling must be sent to the data logger, which can be devices such as microcontroller, microcomputer, etc. In the issue of recording the acceleration during the passage of the shock wave, the speed of data transmission is very important to avoid loss of measurement data. An example of a communication interface that is straightforward to implement and possible to use in the issue under consideration is the serial peripheral interface (SPI). This interface is most often used for systems requiring low and medium data transfer rates (Anand *et al.*, 2014), but can also be successfully used in systems requiring higher communication speeds (Brezeanu *et al.*, 2022; Coşkun *et al.*, 2023; Mohd Noor & Saparon, 2012). An important element of the communication system that needs to be considered is also the distance over which the communication is carried out. In contrast to long-distance communication protocols such as USB (Universal Serial Bus), PCI (Peripheral Component Interconnect), and Ethernet (Park & Mackay, 2003), SPI along with interfaces such as I2C (Inter-Integrated Circuits) and CAN (Control Area Network), is commonly used for short and medium-distance communication. In the issue of recording shock wave parameters, long-distance communication is most often required for safety reasons, which involves signal conditioning, especially for low-voltage signals like those from MEMS-type accelerometers. Signal conditioning is associated with an increase in interference, so the use of communications like SPI with a properly protected data logger could be an improvement in this type of measurement. SPI, compared to other similar communication protocols, is characterized by high transmission rates combined with minimal hardware requirements (Vijaya *et al.*, 2011). SPI has a higher transmission speed than similar communication protocols like I2C and UART (Universal Asynchronous Receiver and Transmitter). In contrast to the semi-duplex nature of I2C, which utilizes a single data line (SDA) and a clock line (SCL), SPI facilitates full-duplex communication through the implementation of dedicated transmit (MOSI – Master Output Slave Input) and receive (MISO – Master Input Slave Output) lines. Conversely, UART, despite its simplicity and frequent utilization for point-to-point communication, also operates in half-duplex (or quasi full-duplex, depending on the implementation) mode and does not offer the ability to exchange data as quickly and simultaneously as SPI. In addition, SPI does not require device addressing like I2C, which simplifies the protocol and allows much higher transmission speeds (even tens of MHz), making it an ideal choice for systems requiring fast, reliable and synchronous communication with multiple devices.

The purpose of this paper is to verify the feasibility of using the SPI communication interface in the design of the authors' MEMS accelerometer-based shock wave feature recorder. The paper discusses technical issues related to the development of a communication interface designed for the problem at hand. Furthermore, a series of experiments has been conducted to ascertain the efficacy of the system under development. As part of this experimental research, a constant voltage was applied to the output of the analog to digital converter (ADC) in order to verify the correct discretization of the input signal and to evaluate the noise of the signal at sampling frequencies up to 45 kHz. In addition, as part of this paper, some comparative research of various measurement systems was carried out on the issue of recording data from dynamic phenomena. As part of this experiment, force and acceleration were measured during the impact of a hammer suspended on a pendulum against a bumper placed at the pendulum's equilibrium center. The conduction of this research enabled a comparison of the quality of recorded data from

different measurement systems, whilst also enabling preliminary verification of the possibility of uprooting the developed recorder based on the MEMS accelerometers in the research of dynamic phenomena. The developed measurement system demonstrates a less common approach to the research of shock wave characteristics, precisely by employing medium-distance communication that does not necessitate the conditioning of the measurement signal. In this approach, the data logger must be relatively close to the sensor, which means that it will be in the range of the shock wave and must be protected from its influence. Furthermore, the development of an original measurement system affords the authors greater autonomy in the development of the software utilized. It is important to note the potential for the utilization of neural networks, which are being employed with increasing frequency in sensory systems (Kciuk *et al.*, 2023), in addressing a particular issue. This involves the development of a model that can ascertain the impact of the wave on human health, with this model being based on the readings obtained from the data logger.

2. Method

The SPI communication interface used in this work operates in a master-slave configuration in half-duplex mode. The chip select (CS) line is responsible for initiating communication between devices. This communication is only active when the line is in a low state, and this state must be maintained for the duration of the communication. It has been established that an oscillating digital clock (DCLK) signal is shared between the Master and the Slave. This signal dictates the timing of bit transmission on the data line (Texas Instruments, 2010). The signal on the DCLK line is generated by the device that is in master communication. The transmission rate is directly related to the clock frequency, since one bit is transmitted only during a single clock cycle. This ensures coordinated data transfer (Lynch *et al.*, 2015) by informing the Master and Slave when to initiate communication. The digital input (DIN) and digital output (DOUT) lines are the data lines telling the Master to send a request to the Slave and sending a response from the Slave to the Master, respectively. Figure 1 shows a simplified diagram of the communication between the microcomputer and the ADC.

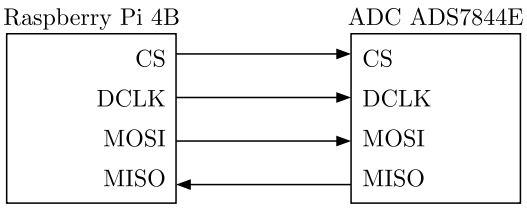

Fig. 1. Simplified diagram of the communication between the registrar (Raspberry Pi 4B) and the ADC.

With reference to the principles of the SPI interface discussed earlier and the characteristics of the transmitter used, the diagram in Fig. 2 shows a detailed sequence of data exchange between the microcontroller and the measurement system. The illustration shows the timing relationships between the basic signals of the SPI bus: DCLK, CS line (CS), input data line (MOSI), and output data line (MISO). Also indicated are the activation moments of the various transmission phases, including communication initialization, transmission of control commands and reading of measurement data. Analysis of this sequence allows fine-tuning of the microcontroller's configuration to meet the timing requirements of the transmitter, which is key to ensuring correct synchronization and reliability of data transmission. The detailed timing and communication scheme for SPI data exchange is shown in Fig. 2.

The measurement system used in the experiment consisted of the following components: Raspberry Pi 4B with Raspberry Pi OS (based on Debian) installed, acting as a control and data logging unit; ADS7844E analog-to-digital converter – a 12-bit, 8-channel chip communicating

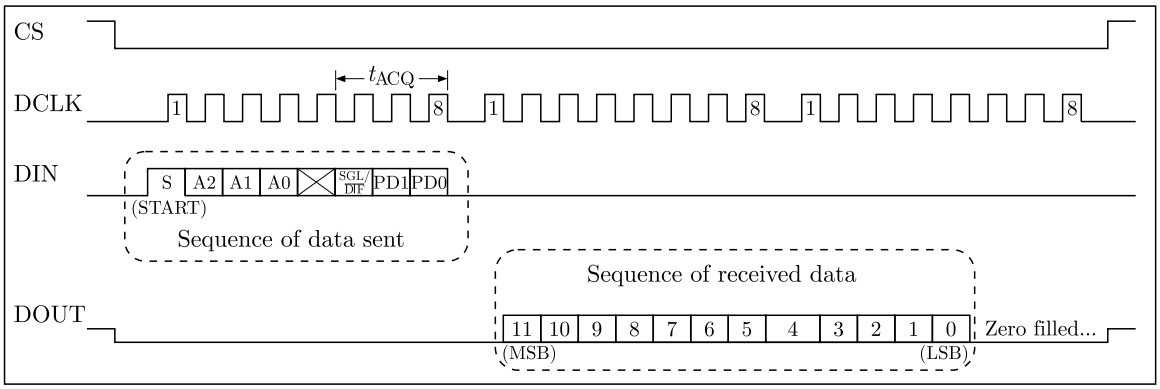

Fig. 2. SPI communication timing diagram as specified in the device datasheet, showing the sequential exchange of control and data bytes.

with the Raspberry Pi via the SPI interface; and a laboratory power supply generating a test voltage in the range of 0 V–3.3 V. The ADS7844E transducer was powered by 3.3 V, supplied by a voltage regulator that also served as the reference voltage source (Vref). SPI bus connections (DCLK, MISO, MOSI, and CS) were established in accordance with the ADS7844E's technical documentation and the Raspberry Pi 4B's GPIO pin layout.

The following waveform (Fig. 3) presents the outcome of measuring SPI communication in "every byte" transmission mode, which was recorded with a digital oscilloscope. Clearly defined sequences of clock pulses (DCLK) and data line activations (MOSI/MISO) can be seen, corresponding to single bytes transmitted in a single transaction. Significantly, the presence of minor interference and noise in the signal lines is observed, which appears only during inactive periods, i.e., between consecutive transmissions. This localization of interference clearly indicates that it is not generated by the MEMS chip itself or the transducer during operation, but is of an external nature (e.g., coming from the power supply, electromagnetic interference or reflections on the transmission lines). This confirms the correctness of the data exchange process itself, while at the same time emphasizing the need to address shielding issues.

Fig. 3. Oscilloscope capture of a single SPI data exchange sequence between the microcontroller and the MEMS sensor, illustrating the timing relationship between clock and data lines.

Figure 4 shows the course of communication in the mode of reading every second byte, used to filter irrelevant information selectively. In this variant, only those fragments of transmission that correspond to the relevant measurement data are received, while bits present outside the main exchange sequence are skipped. As a result, the recorded waveform is clearer, and only key moments related to the transmission of utility values are analyzed.

In order to reduce distortion and minimize the impact of noise on the measurement signal, the data acquisition process uses a strategy of receiving every second response from the SPI

Fig. 4. Oscilloscope capture of SPI communication in reduced sampling mode (every second byte), showing selective acquisition of relevant data segments.

system. The received data is then processed using a bit mask, taking into account the fact that the oldest bit (MSB) is transmitted first. If an interference occurs when the most significant bit is transmitted, the potential error is 50 % of the full measurement range. The designed algorithm effectively bypasses this kind of failure, providing transmission error robustness. In addition, the use of a bit mask makes it possible to correct the data already at the recording stage, which significantly increases the reliability of the system. Despite the simplified filtering procedure, the solution makes it possible to achieve a stable sampling frequency of 45 kHz, which is a significant advantage over piezoelectric sensor systems, which achieve about 20 kHz per channel. The algorithm for processing a single packet of data is presented in Fig. 5.

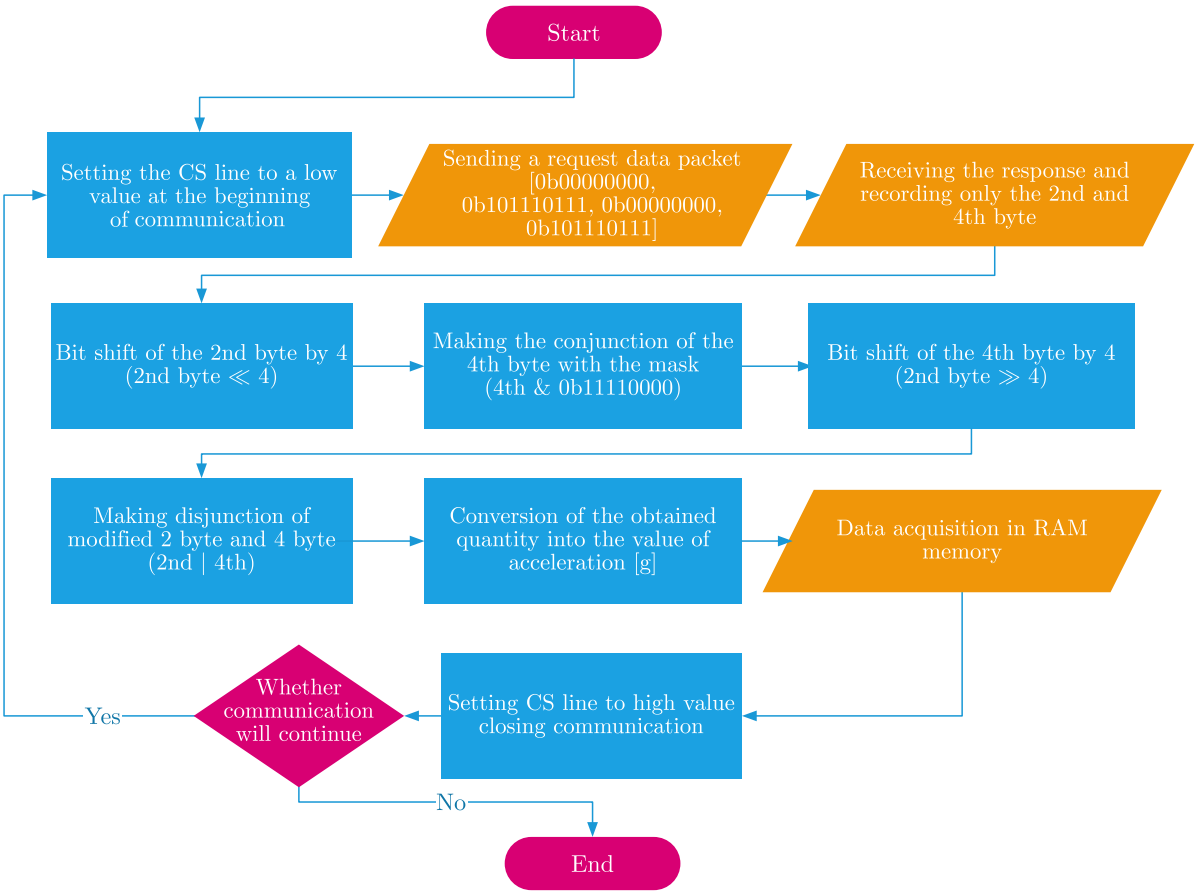

Fig. 5. Block diagram of the algorithm for data transmission, reception, and processing within the MEMS-based measurement system.

As a preliminary test, the DC voltage was measured from a bench power supply, applied to one of the analog inputs of the ADS7844E converter. The purpose of the test was to verify the correct operation of the entire measurement path and to evaluate the accuracy of the conversion of the analog signal to digital form. The output voltage from the power supply was set at several levels (0.6 V, 1.65 V, 2.4 V, 3.3 V) – including those corresponding to the power supply of the sensor used later in the project. Data reading from the transmitter was implemented using the Python language and the spidev library to handle the SPI bus. The raw digital values were scaled to the corresponding voltage values, and then compared with measurements taken in parallel using a multimeter to assess the accuracy of the transmitter.

Figure 6 shows the voltage readings measured by the ADS7844E converter. The readings from the converter show good agreement with the reference values, which confirms the correctness of the measurement path. The voltages are stable, and the differences were within the acceptable range of measurement error (oscillations due to power supply interference). The initial and final distortions seen in the graph are due to signal filtering effects associated with the limited sampling range and the operation of the smoothing algorithm. These results confirm that the circuit can be used for further measurements in the project.

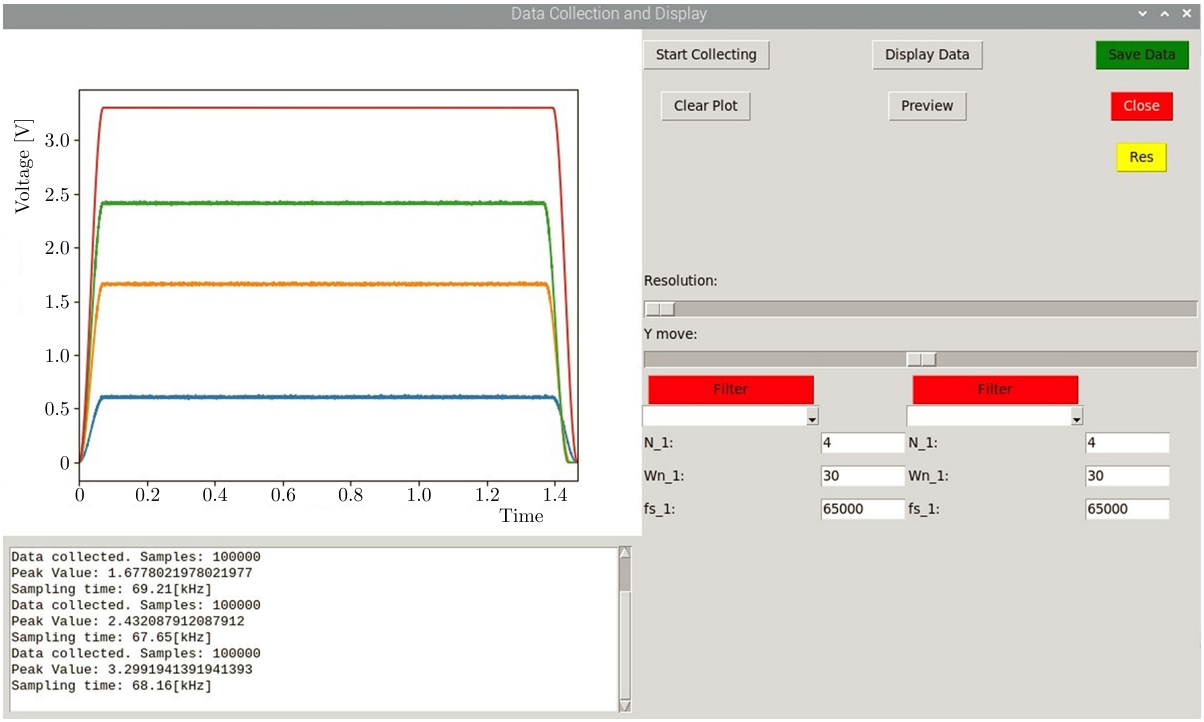

Fig. 6. Graphical user interface (GUI) screenshot illustrating the test of SPI communication between the ADC module and the microcomputer.

A physical pendulum is any rigid body that can make free oscillatory motions about a horizontal axis not passing through its center of mass, under the influence of gravity. Unlike the ideal mathematical pendulum – which is a model of a material point suspended from a weightless and inextensible thread – the physical pendulum takes into account the actual geometric properties and mass distribution of the body. In the experiment conducted, this system was used to generate a controlled force pulse by striking a force sensor with the tip of the pendulum. At the same time, the response of the measurement system based on a MEMS accelerometer integrated into a Raspberry Pi 4B microcomputer via an SPI interface was recorded. The results were compared with values obtained from a piezoelectric sensor system, built with independent acceleration sensors, to evaluate the accuracy and sensitivity of the MEMS system under dynamic forcing conditions.

During the experiment, the pendulum shown in Fig. 7 was released from two fixed angles of deflection: 20° and 40°. At the lowest point of the trajectory, contact was made between the tip of the pendulum and the FC500 force sensor (AXIS Sp. z o.o., Gdansk, Poland) (FC500), allowing precise measurement of the impact force. Simultaneous data recording was carried out from three independent measurement systems: the force sensor, a MEMS ADXL377 accelerometer (Analog Devices, USA) (ADXL377) connected to a Raspberry Pi 4B microcomputer via an SPI interface, and a piezoelectric sensor system – the MTS DSP SigLab 20-42 Dynamic Signal Analyzer (MTS Systems Corporation, USA) – consisting of PCB Piezotronics 333B31 (PCB Piezotronics, USA) (333B31) and 352C33 (PCB Piezotronics, USA) (352C33) acceleration sensors. Due to the temporal synchronization of all measurement channels, it was possible to compare the obtained results and assess their mutual compatibility directly. The purpose of the measurements was both to calibrate the SPI-based data acquisition system and to verify the quality of the MEMS accelerometer signal readout in the context of future applications for analyzing dynamic phenomena. For each of the two tilt angles, 30 independent measurements were made, ensuring the statistical reliability of the data obtained.

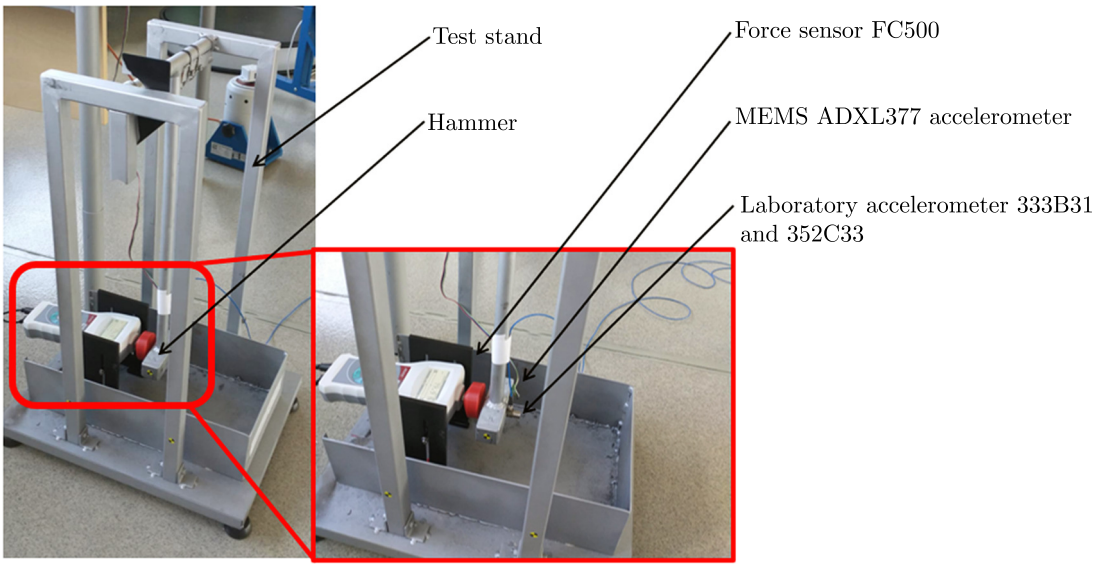

Fig. 7. Experimental setup with physical pendulum and mounted sensors, including a force sensor at the impact point, a MEMS accelerometer, and reference acceleration sensors. The MEMS unit is connected to a Raspberry Pi 4B for acquisition of dynamic response data via SPI interface.

3. Results and discussion

Experimental research was performed for two selected pendulum swing angles: 20° and 40°. In each case, the pendulum was released from a fixed angle and then struck a force sensor fixed at a fixed point. Based on the recorded signal from the sensor, the maximum value of the peak impact force was determined. At the same time, accelerations obtained from two independent sources were recorded: from a MEMS accelerometer integrated into a Raspberry Pi-based measurement system and from piezoelectric acceleration sensors, which are a separate measurement system (MTS DSP SigLab). The results of 30 consecutive impacts are summarized for the 20° and 40° deflection accounts in Table 1.

The lowest recorded spread of values over a period of 30 consecutive measurements has been documented for the FC500 force sensor, as evidenced by the coefficient of variation, which does not exceed 3 %. Additionally, the low spread of data is characterized by results recorded with the use of a logger, employing a MEMS capacitive accelerometer, for which the coefficient of variation does not exceed 5 %. Data recorded with the MEMS accelerometer is characterized by the

Table 1. Maximum values of the recorded force (FC500 force sensor) and acceleration (MEMS accelerometer, piezoelectric accelerometers 333B31 and 352C33) during 30 consecutive hammer impacts for the pendulum swing angle of 20° and 40°.

	FC500		MEMS		333B31		352C33	
Swing angle	20°	40°	20°	40°	20°	40°	20°	40°
	72.60	155.50	15.17	28.85	12.80	24.10	10.70	23.30
	75.30	156.50	14.39	28.46	12.70	23.80	10.40	23.80
	79.90	150.20	15.76	28.85	15.20	27.40	14.40	24.90
	78.60	151.80	16.15	29.82	14.30	29.90	14.60	29.00
	76.60	152.70	14.78	29.24	14.70	29.80	14.00	29.50
	75.10	151.80	14.98	28.85	14.50	29.10	14.70	28.80
	74.00	153.80	14.98	29.24	12.90	30.00	12.60	28.60
	78.40	155.00	16.54	30.02	14.20	30.70	12.10	28.80
	74.70	154.90	14.78	29.63	15.00	30.50	14.80	29.50
	73.30	154.20	14.00	29.82	11.10	30.00	9.50	29.50
	75.00	155.10	14.98	30.61	14.60	30.70	14.50	29.50
	77.00	153.30	14.98	30.00	14.40	29.00	12.60	27.70
	74.00	153.80	14.59	30.02	11.10	29.50	8.60	28.80
	77.10	151.60	16.74	29.24	15.50	29.00	15.10	27.90
	75.90	154.40	14.78	30.41	14.70	29.80	14.30	28.30
	74.00	153.00	15.37	29.24	13.00	29.40	11.00	28.00
	74.00	153.60	14.78	29.24	14.70	29.10	14.50	25.70
	76.10	154.20	14.78	29.63	14.40	28.60	12.70	25.60
	77.30	154.80	15.37	30.41	12.00	29.50	10.80	27.50
	77.20	154.30	15.95	30.02	11.20	29.40	9.00	26.20
	77.40	154.30	14.98	29.63	13.90	29.40	11.70	26.10
	82.50	155.30	16.15	30.22	15.40	29.90	15.50	26.60
	75.40	153.40	15.37	30.41	15.00	29.40	15.00	26.30
	77.30	154.40	15.95	30.60	11.00	29.60	8.90	26.10
	75.20	155.10	14.39	30.20	15.50	28.40	15.00	26.20
	75.80	153.40	14.78	30.41	11.30	28.30	8.70	27.10
	76.30	154.20	14.78	30.20	16.00	27.60	15.20	25.40
	75.60	154.30	15.95	29.20	11.30	27.50	10.00	25.50
	75.60	154.70	14.78	30.41	15.50	27.20	14.90	23.60
	76.10	155.40	15.17	30.02	15.60	28.00	15.30	25.20
$\overline{x}$	76.11	153.97	15.21	29.76	13.78	28.82	12.70	26.97
σ	2.03	1.34	0.67	0.60	1.63	1.62	2.37	1.86
CV	2.66 %	0.87 %	4.38 %	2.01 %	11.85 %	5.64 %	18.68 %	6.91 %

lowest standard deviation among all acceleration meters, even though the average peak acceleration value for MEMS was the highest. The acceleration recorded by piezoelectric accelerometers is characterized by the largest data spread. At the same time, the spread of data for piezoelectric accelerometers is significant, and for the 352C33 sensor, the coefficient of variation reaches almost 20 %. This is caused by the frequent repetition for this dataset of measured values that are much smaller than the others and much smaller than the values recorded by the MEMS accelerometer. The average value measured by the 352C33 sensor is 12.70 g, with results as low

as less than 9 g being repeated among the recorded data, i.e., values that are far from the average value by more than the standard deviation. For hammer impacts released at a pendulum swing angle of 40°, the values recorded by all sensors increased in proportion to the increase in the value of the pendulum swing angle. The spread of data between successive measurement systems exhibits analogous characteristics to those observed at 20°. Again, the lowest coefficient of variation is characterized by the data set recorded by the force sensor (less than 1 %), and a comparable level of spread is observed for MEMS (slightly above 2 %). As for the first dataset, the greatest spread in the data is characterized by the results recorded by piezoelectric accelerometers. However, for measurements at a pendulum swing angle of 40°, the recorded spread is much smaller and for piezoelectric accelerometers does not exceed 7 %. The decrease in spread for all measurement systems at impacts for a pendulum swing angle of 40° is most likely due to the fact that, for this case, we observe a larger change in the measured value, which makes it easier to register by the used measurement systems. In order to better show the spread of the data for the individual accelerometers, Fig. 8 shows box plots for impacts with pendulum swing angles of 20° and 40°.

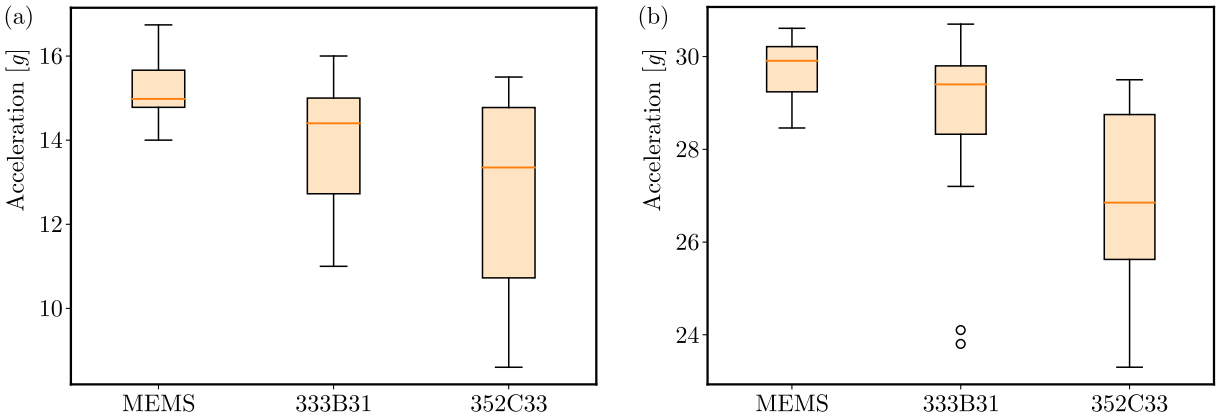

Fig. 8. Box plot showing the comparison of the measurement results of individual devices at impact for: (a) 20°; (b) 40°.

Box plots provide a graphical representation of the features of the data sets being ranked. Based on the box plots shown, it can be seen that the values recorded for the MEMS accelerometer are higher than for the other sensors, as evidenced by the median value, which is highest for the MEMS sensor. The difference in magnitude is most likely due to the fact that the recorder using the MEMS accelerometer records the data at the highest sampling rate, which significantly increases the chances of recording key points during the course of the impact. It can be seen that in all graphs the median is shifted relative to the center of the corresponding box plots, which indicates the asymmetry of the recorded data. In order to provide a more accurate illustration of the characteristics of the recorded data distributions, Figs. 9 and 10 present histograms for the data that was recorded at 20° and 40°, respectively.

As illustrated in the histograms, the normal distribution curve has been superimposed for the purpose of evaluating the resulting distributions, with the said curve having been drawn for the corresponding data sets based on their mean and standard deviation. The graphs presented in Fig. 9 show the asymmetric nature of the distribution for all measuring instruments. However, for the FC500 force sensor and for the MEMS accelerometer, the data sets manifest some characteristics of a normal distribution, and it can be assumed that with an increase in the measurement sample, the results for these sensors could be estimated with the help of a normal distribution. However, the results for piezoelectric accelerometers are characterized by a large asymmetry. Furthermore, a high frequency of occurrences is observed in the upper and lower extreme intervals of the histogram. This is an unexpected result for this type of measurement, and it is not recorded by other measurement systems. This is most likely related to the frequent failure of piezoelectric sensors to register the maximum value in the peak. This

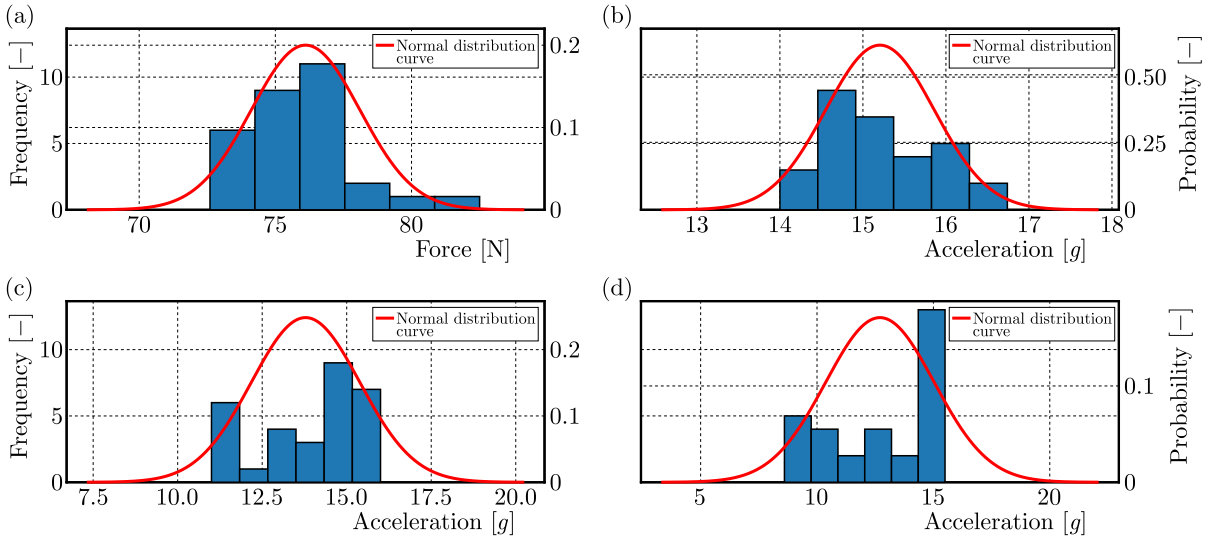

Fig. 9. Histograms showing the distribution of measurement results for different devices with a normal distribution curve determined from the sample mean and standard deviation at impact for 20°. Histograms are shown successively for: (a) FC500 force sensor; (b) MEMS-type accelerometer; (c) 333B31 piezoelectric sensor; (d) 352C33 piezoelectric sensor.

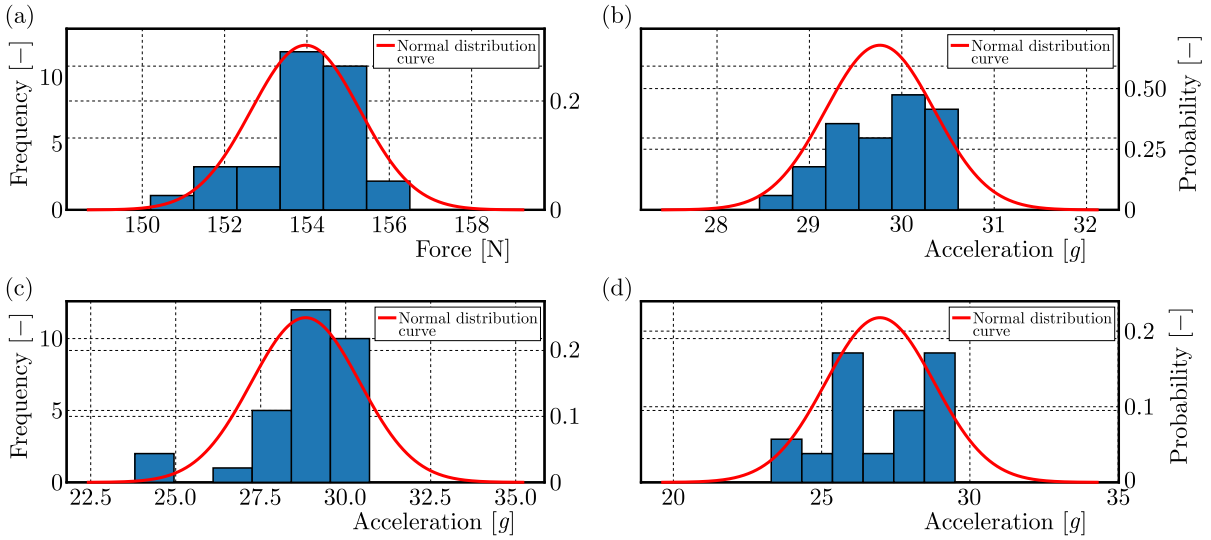

Fig. 10. Histograms showing the distribution of measurement results for different devices with a normal distribution curve determined from the sample mean and standard deviation at impact for 40°. Histograms are shown successively for: (a) FC500 force sensor; (b) MEMS-type accelerometer; (c) 333B31 piezoelectric sensor; (d) 352C33 piezoelectric sensor.

leads to registering, for the maximum acceleration value, the value occurring before or after the maximum value in the peak, and therefore, registering a smaller value than the actual value.

The distribution of rasterized data for an impact at a pendulum deflection angle of 40° is analogous to that for 20°. It is important to note that the data recorded by the 333B31 sensor is an exception to this. Indeed, for this particular angle, the data shows the characteristics of a normal distribution to a higher degree. However, it is also important to note that significant outliers are observed here, as evidenced by a break in the continuity of the histogram. In the histograms, individual results can be observed that significantly deviate from the other measured values (which is particularly evident for the 333B31 piezoelectric accelerometer). If the purpose of the conducted research was to determine acceleration values, these results should be discarded.

However, the purpose of the conducted research is to compare different measurement methods. The frequency of occurrence of this type of interference is an important element from the point of view of this comparison. In addition, it testifies in favor of the assumption that for piezoelectric accelerometers, the observed inaccuracies are the result of not recording the maximum value of acceleration in the peak.

Also important from the point of view of evaluating the characteristics of the shock wave is the quality of recording the acceleration waveform. To visualize this, Fig. 11 shows the acceleration waveform during the impact of the hammer with a pendulum swing of 20°.

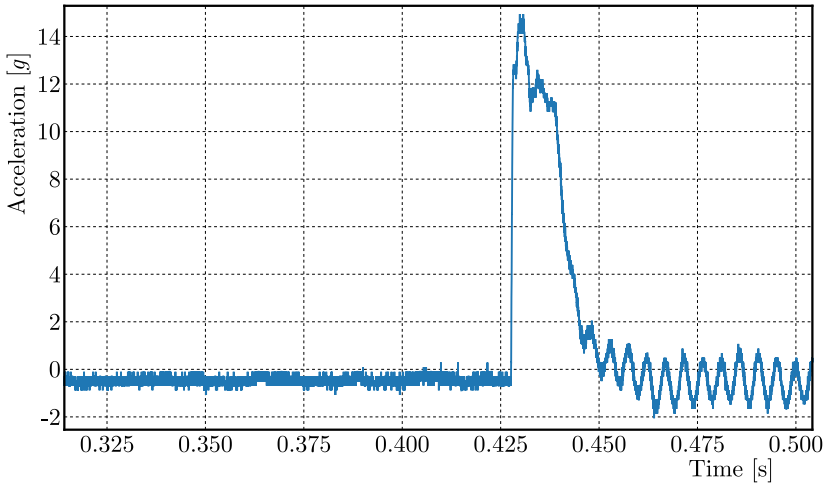

Fig. 11. Waveform of acceleration during the impact of the hammer at a pendulum deflection of 20°.

It can be observed that the authors' data logger managed to record the acceleration peak occurring during the hammer impact to a degree that allowed determining the maximum value and also determining the time of impact. In addition, using the MEMS accelerometer, it was also possible to record vibrations occurring in the system immediately after the hammer impact, which is visible in the form of a harmonic acceleration course visible immediately after the impact.

4. Conclusions

As a result of the conducted research, it can be concluded that the use of a system based on the SPI protocol with a MEMS sensor allows precise and repeatable results to be obtained in various dynamic applications. Comparison of the results with the piezoelectric sensor system based on force sensors and piezoelectric accelerometers confirms the consistency of measurements and the effectiveness of using MEMS in the analysis of dynamic phenomena.

Conclusions:

- universality of the SPI system: the system based on the SPI protocol with a MEMS sensor provides flexibility and wide application possibilities in dynamic measurements;
- wider frequency range: the SPI system with MEMS works correctly up to 45 kHz, which gives an advantage over traditional piezoelectric sensor systems that only work up to 20 kHz;
- repeatability of measurements: the system demonstrates high repeatability of results and low scatter, which confirms the reliability of the MEMS sensor in measurements;
- compatibility with piezoelectric sensors system: the results obtained from the MEMS sensor-based SPI system are consistent with those of the piezoelectric sensors system, which demonstrates its precision.

References

1. ADXL377. Retrieved January 10, 2025, from: https://www.analog.com/en/products/adxl377.html
2. Anand, N., Joseph, G., Oommen, S.S., & Dhanabal, R. (2014). Design and implementation of a high speed Serial Peripheral Interface. *2014 International Conference on Advances in Electrical Engineering (ICAEE). IEEE.* https://doi.org/10.1109/ICAEE.2014.6838431
3. Arkusz, K., Klekiel, T., Sławiński, G., & Będziński, R. (2019). Influence of energy absorbers on Malgaigne fracture mechanism in lumbar-pelvic system under vertical impact load. *Computer Methods in Biomechanics and Biomedical Engineering*, *22*(3), 313–323. https://doi.org/10.1080/10255842.2018.1553238
4. Baranowski, P., Małachowski, J., & Mazurkiewicz, Ł. (2020). Local blast wave interaction with tire structure. *Defence Technology*, *16*(3), 520–529. https://doi.org/10.1016/j.dt.2019.07.021
5. Brezeanu, I.B., Botezatu, C., Drăghici, F., & Brezeanu, G. (2022). Improved SPI controlled, low-voltage, high speed, multi-channel switch. *2022 14th International Conference on Electronics, Computers and Artificial Intelligence (ECAI). IEEE.* https://doi.org/10.1109/ECAI54874.2022.9847444
6. Coşkun, O., Egeli, E., Tarcan, E., Kurtuluş, İ., & Yılmaz, G. (2023). Analysis and implementation of a daisy-chain serial peripheral interface bus for a communication network with multiple PLC modules. *2023 14th International Conference on Electrical and Electronics Engineering (ELECO). IEEE.* https://doi.org/10.1109/ELECO60389.2023.10416045
7. Covaciu, D., & Dima, D.S. (2017). Crash tests data acquisition and processing. In A. Chiru, & N. Ispas (Eds.), *CONAT 2016 International Congress of Automotive and Transport Engineering* (pp. 782–789). Springer, Cham. https://doi.org/10.1007/978-3-319-45447-4_85
8. Dima, D.S., & Covaciu, D. (2017). Solutions for acceleration measurement in vehicle crash tests. *IOP Conference Series: Materials Science and Engineering*, *252*, Article 012007. https://doi.org/10.1088/1757-899X/252/1/012007
9. Elsayed, N.M., & Atkins, J.L. (Eds.). (2008). *Explosion and blast-related injuries: Effects of explosion and blast from military operations and acts of terrorism.* Elsevier Academic Press.
10. Erdik, A., Kilic, S.A., Kilic, N., & Bedir, S. (2016). Erratum to: Numerical simulation of armored vehicles subjected to undercarriage landmine blasts (Shock Waves, 10.1007/s00193-015-0576-1). *Shock Waves*, *26*(4), 531. https://doi.org/10.1007/s00193-016-0678-4
11. FC500. Retrieved January 10, 2025, from: https://www.axis.pl/en/fc/405-fc500.html
12. Kciuk, M., Kowalik, Z., Lo Sciuto, G., Sławski, S., & Mastrostefano, S. (2023). Intelligent medical velostat pressure sensor mat based on artificial neural network and Arduino embedded system. *Applied System Innovation*, *6*(5), Article 84. https://doi.org/10.3390/asi6050084
13. Kciuk, S., Krzystała, E., Mężyk, A., & Szmidt, P. (2022). The application of microelectromechanical systems (MEMS) accelerometers to the assessment of blast threat to armored vehicle crew. *Sensors*, *22*(1), Article 316. https://doi.org/10.3390/s22010316
14. Lynch, K.M., Marchuk, N., & Elwin, M.L. (2015). *Embedded computing and mechatronics with the PIC32 microcontroller.* Newnes.
15. Mohd Noor, N.B., & Saparon, A. (2012). FPGA implementation of high speed serial peripheral interface for motion controller. *2012 IEEE Symposium on Industrial Electronics and Applications.* IEEE. https://doi.org/10.1109/ISIEA.2012.6496676
16. Olvey, S.E., Knox, T., & Cohn, K.A. (2004). The development of a method to measure head acceleration and motion in high-impact crashes. *Neurosurgery*, *54*(3), 672–677. https://doi.org/10.1227/01.NEU.0000108782.68099.29
17. Park, J., & Mackay, S. (2003). Appendix F – Number systems. In *Practical Data Acquisition for Instrumentation and Control Systems* (pp. 389–397). Elsevier. https://doi.org/10.1016/B978-075065796-9/50018-5

18. Pyka, D., Kurzawa, A., Żochowski, P., Bajkowski, M., Magier, M., Grygoruk, R., Roszak, M., Jamroziak, K., & Bocian, M. (2025). Experimental and numerical research on additional vehicles protection against explosives. *Archives of Civil and Mechanical Engineering*, *25*(2), Article 83. https://doi.org/10.1007/s43452-025-01121-w

19. Sethuramalingam, T.K., & Vimalajuliet, A. (2010). Design of MEMS based capacitive accelerometer. *2010 International Conference on Mechanical and Electrical Technology*. IEEE. https://doi.org/10.1109/ICMET.2010.5598424

20. Texas Instruments (2010, rev. 2019). KeyStone Architecture Serial RapidIO (SRIO). User's Guide. https://www.ti.com/lit/ug/sprugw1c/sprugw1c.pdf

21. Vijaya, V., Valupadasu, R., Chunduri, B.R., Rekha, C.K., & Sreedevi, B. (2011). FPGA implementation of RS232 to Universal serial bus converter. *2011 IEEE Symposium on Computers and Informatics*. IEEE. https://doi.org/10.1109/ISCI.2011.5958920

22. Wrazidło, D., Sławski, S., Krzystała, E., & Jarosz, T. (2025). Numerical analysis of shock wave impact on a gas cylinder. In E. Świtoński, A. Mężyk, S. Kciuk, & R. Szewczyk (Eds.), *Lecture Notes in Networks and Systems: Vol. 1146. PCM—CMM2023: Theories, Models and Simulations of Complex Physical Systems* (pp. 196–207). Springer. https://doi.org/10.1007/978-3-031-73161-7_18

23. 333B31. Retrieved January 10, 2025, from: https://www.pcb.com/products?m=333b31

24. 352C33. Retrieved January 10, 2025, from: https://www.pcb.com/products?m=352c33

Manuscript received May 5, 2025; accepted for publication June 18, 2025; published online October 7, 2025.

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 769–780, Warsaw 2025
https://doi.org/10.15632/jtam-pl/208127

THE MODEL-BASED ALGORITHM FOR AUTONOMOUS VEHICLE PATH FOLLOWING

Michał BRZOZOWSKI

Department of Combustion Engines and Vehicles, University of Bielsko-Biala, Bielsko-Biala, Poland
mbrzozowski@ubb.edu.pl

This paper presents a proprietary steering algorithm for path following, whose advantage lies in its ability to be applied with vehicle dynamic models of varying complexity. The proposed algorithm was implemented using models with 3 and 10 degrees of freedom, which had been previously verified. The results were compared with those obtained using the geometric pure pursuit (PP) algorithm. Both algorithms require path approximation. In this study, path approximation was conducted using B3 functions. The presented computer simulation results indicate that the proposed steering angle selection algorithm demonstrates greater accuracy than the PP algorithm.

Keywords: autonomous vehicle; vehicle dynamics; vehicle modelling.

1. Introduction

One of the more challenging problems in the deployment of highly automated vehicles is the provision of an appropriate control system that enables the execution of a predefined driving path. This task involves determining the steering angle trajectory of the front wheels based on a known velocity profile in such a way that the vehicle follows the desired path. Numerous vehicle control methods exist, and their comparisons and benchmarking can be found in works such as Liu *et al.* (2021) and Diachuk and Easa (2022). Path-following methods generally fall into three main categories: geometric methods, model-based methods, and learning-based (artificial intelligence) methods.

Among the popular geometric methods, which rely on the curvature of the path, the pure pursuit (PP) algorithm stands out. Originally formulated in the 1990s (Coulter, 1992), it is still widely used in autonomous driving applications (Gámez-Serna *et al.*, 2017). Huang *et al.* (2018) present an application of the PP algorithm, focusing on the determination of the constant l_d, which is essential for the proper functioning of the algorithm and is related to vehicle speed. Setting the constant too high reduces trajectory-tracking accuracy, which becomes critical during maneuvers such as U-turns. Conversely, a too-low value leads to oscillations in the steering angle. This algorithm continues to be used in vehicle control tasks, for example in (An *et al.*, 2025).

Another well-known geometric control algorithm is Stanley control (SC), whose application was described by Yang *et al.* (2017), with particular attention to issues related to path curvature computation, which significantly influences the behavior of geometric algorithms. An alternative

Ministry of Science and Higher Education
Republic of Poland

This publication has been funded by the Polish Ministry of Science and Higher Education under the Excellent Science II programme "Support for scientific conferences".
The content of this article was presented during the 61st Symposium "Modelowanie w mechanice" (Modelling in Mechanics), Szczyrk, Poland, March 2–5, 2025.

use of the SC algorithm was proposed by Amer *et al.* (2018), where it was combined with optimization methods. Cibooglu *et al.* (2017) present a hybrid approach combining the PP and SC algorithms to enhance precision. Other, less common geometric algorithms include the one described in the monograph by Rajamani (2012). A comprehensive comparison of four geometric algorithms is provided by Brzozowski (2025).

The advantages of geometric algorithms include their ease of implementation, low computational requirements, and high effectiveness in simple scenarios (e.g., low-curvature routes). Their disadvantages include poor performance at higher speeds, moderate trajectory-tracking accuracy, and sensitivity to the tuning of auxiliary/scaling parameters. Additionally, these methods do not account for the vehicle dynamics model.

Model-based methods (using either kinematic or dynamic vehicle models) describe the vehicle's motion. Typical examples include optimization-based approaches such as model predictive control (MPC) and the linear quadratic regulator (LQR). These methods are also referred to as optimal control strategies.

The LQR method uses a feedback gain matrix to minimize a cost function (Li *et al.*, 2019; Lee *et al.*, 2019). While LQR provides greater accuracy than geometric algorithms, it is less computationally efficient. A significant drawback is the necessity of linearization and the inability to handle constraints.

Another class of optimization-based methods is represented by MPC, which predicts the future behavior of the vehicle over a sequence of subintervals within the total planning horizon based on a mathematical model. MPC performs real-time optimization to follow a reference path while satisfying constraints. Variants of this model are discussed in (Zhang *et al.*, 2019) and (Fu *et al.*, 2022), where the algorithm is adapted to specific needs. MPC's advantages include its ability to incorporate constraints, predict and avoid problems in advance, and work with nonlinear vehicle models. It is considered a highly accurate control method, although its disadvantages include high complexity and low computational efficiency.

A large class of control approaches consists of learning-based methods. These do not require the development of an explicit vehicle model and are suitable when detailed information about the vehicle or its environment is unavailable, but large amounts of driving data exist. MPC is often used as a preliminary tool for training such models. An example of reinforcement learning–based control can be found in (Cao *et al.*, 2023). Among learning-based methods, fuzzy logic–based approaches are also noteworthy, such as in (Elsayed *et al.*, 2018).

In this study, a proprietary algorithm was applied to solve the path-following task. Although it does not directly fall into any of the previously mentioned categories, it requires a dynamic vehicle model for its operation. Therefore, it can be classified alongside methods such as LQR and MPC. It is assumed that the equations of motion take the following form:

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} = \mathbf{f}\left(\mathbf{q}, \dot{\mathbf{q}}, F(t), u\right), \tag{1.1}$$

where $\mathbf{M}$ – inertia matrix, $\mathbf{q}$ – vector of generalized coordinates, u – control parameter.

The linearization of the system would consist in transforming these equations into the following form:

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} = \mathbf{h}(\mathbf{q}, \dot{\mathbf{q}}, F(t)) + \mathbf{G}u, \tag{1.2}$$

where $\mathbf{G}$ – input (or control) distribution matrix, $\mathbf{h}$ – vector of generalized forces.

In the case where the control parameter is $u = \delta$ (the steering angle of the front wheels), it can be determined in one of the following ways:

- as the result of an optimization process (as in LQR, NLQR, MPC, or NMPC methods);
- as a solution to the problem of selecting the steering angle δ by repeated integration of the equations of motion, as proposed in this study. This method does not require linearization.

The algorithm proposed in this study (referred to as MPC/B) is an approach that enables the determination of the steering angle δ, using a vehicle model of arbitrary complexity, without the need for linearization or the application of optimization methods. Control inputs are determined in stages – within sequentially occurring subintervals of the total simulation time. To evaluate the proposed control algorithm, simulation studies were conducted, comparing the results obtained using the MPC/B algorithm to those achieved with the PP algorithm. Two vehicle dynamic models were formulated, and the path was approximated using third-degree spline functions.

Although the geometric PP algorithm does not require a dynamic model to compute the steering angle, it does require information about the vehicle's position in space. In the case of simulation studies, this positional information is provided by the vehicle dynamics model.

2. Nonlinear vehicle models

This study presents two vehicle models with 3 and 10 degrees of freedom (DoF). The 10-DoF model is a three-dimensional model. It is formulated under the assumption that the vehicle is treated as a rigid body with 6 degrees of freedom (the chassis), to which four rotating wheels are attached (Fig. 1).

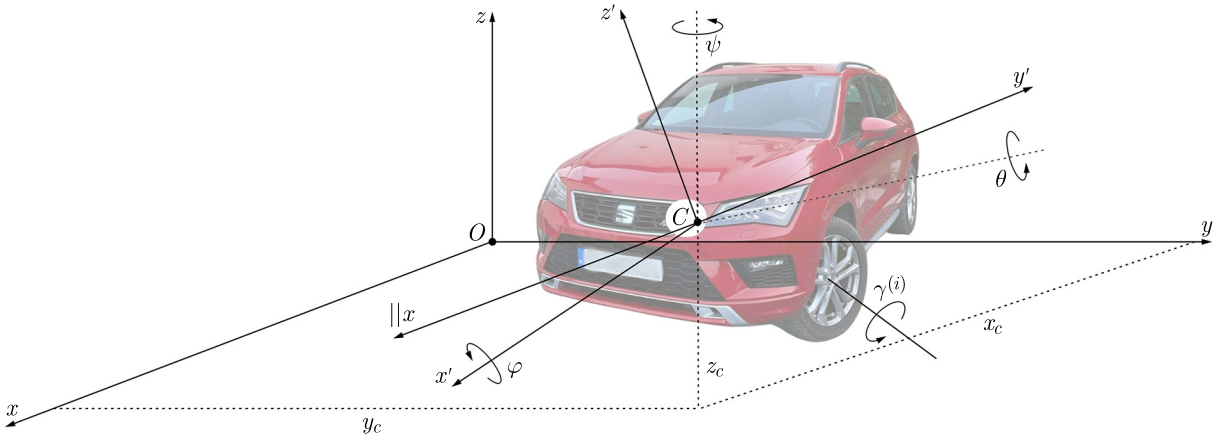

Fig. 1. Diagram of the 10 degrees of freedom model.

To derive the equations of motion for the chassis along with the attached concentrated masses (including wheels and suspensions), a formalism based on the Newton-Euler equations was applied (Blajer, 1998). The Newton-Euler equations for the chassis take the following form:

$$m\dot{\mathbf{V}}_c = \sum_i \mathbf{F}_i, \qquad \frac{\mathrm{d}\mathbf{k}_c}{\mathrm{d}t} = \sum_i \mathbf{M}_{i_c}, \tag{2.1}$$

where $\mathbf{V}_c$ – the velocity vector of the body center of mass, $\mathbf{k}_c$ – angular velocity of the body relative to the center of mass C, $\mathbf{F}_i$ – external forces acting on the body, $\mathbf{M}_{i_c}$ – moments of external forces relative to the center of mass C, m – total mass of the vehicle including the wheels.

The equations of motion for the wheels can be written in the form:

$$I^{(i)}\ddot{\gamma}^{(i)} = \sum_j M_j'^{(i)}, \quad i = 1, 2, 3, 4, \tag{2.2}$$

where $I^{(i)}$ – moment of inertia of the wheel about its axis of rotation, $\gamma^{(i)}$ – wheel rotation angle, $\sum_j M_j'^{(i)}$ – sum of moments of forces acting on the wheel about its axis of rotation.

The vector of generalized coordinates comprising the body and the four wheels of the vehicle thus takes the form:

$$\mathbf{q} = \begin{bmatrix} \mathbf{q}_n \\ \boldsymbol{\Gamma} \end{bmatrix} = \begin{bmatrix} \mathbf{r}_c \\ \boldsymbol{\Phi} \\ \boldsymbol{\Gamma} \end{bmatrix}, \tag{2.3}$$

where

$$\mathbf{r}_c = \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = \begin{bmatrix} q_1 \\ q_2 \\ q_3 \end{bmatrix}, \qquad \boldsymbol{\Phi} = \begin{bmatrix} \varphi \\ \theta \\ \psi \end{bmatrix} = \begin{bmatrix} q_4 \\ q_5 \\ q_6 \end{bmatrix}, \qquad \boldsymbol{\Gamma} = \begin{bmatrix} \gamma^{(1)} \\ \gamma^{(2)} \\ \gamma^{(3)} \\ \gamma^{(4)} \end{bmatrix} = \begin{bmatrix} q_7 \\ q_8 \\ q_9 \\ q_{10} \end{bmatrix}.$$

To determine the road reaction forces on the wheels, the brush model (Pacejka & Sharp, 1991), modified and described in detail in (Rajamani, 2012), was applied. A detailed description of the 10 degrees of freedom dynamic model and the tire model used can be found in (Brzozowski, 2025). In vehicle dynamics modeling for autonomous driving, the most commonly used dynamic model is the planar model with 3 degrees of freedom, also known as the bicycle or moped model (Gillespie, 1992; Ajanović *et al.*, 2023). Figure 2 shows the vehicle representation in the 3 degrees of freedom model. This model accounts for the vehicle's displacement in the xy plane and the yaw angle ψ around z'-axis of the local coordinate system $\{C\}'$ which is parallel to the z-axis of the road coordinate system $\{O\}$. The vehicle dynamics are described by the components of the vector shown in Fig. 2:

$$\mathbf{q} = \begin{bmatrix} V'_x \\ V'_y \\ \psi \end{bmatrix}, \tag{2.4}$$

where V'_x, V'_y – the components of the vehicle velocity vector in the local coordinate system $\{C\}'$, ψ – yaw angle.

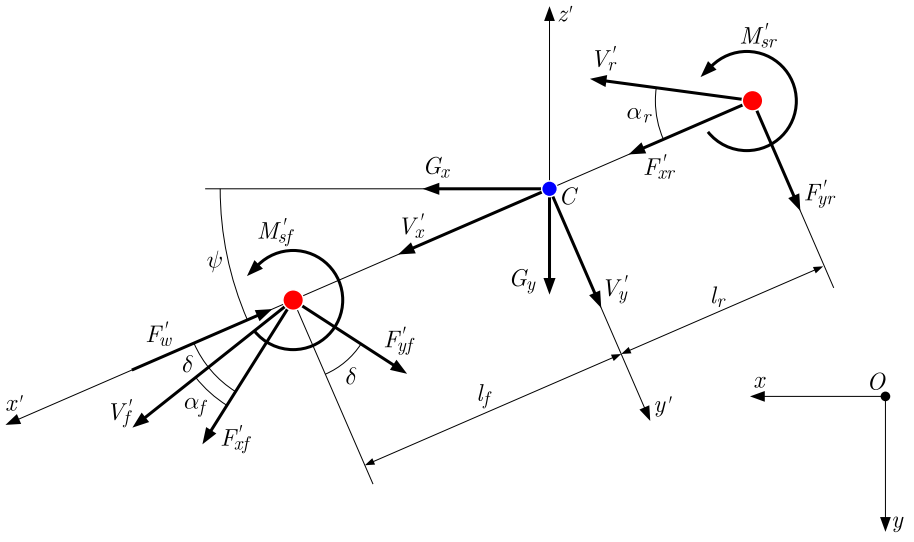

Fig. 2. Diagram of the bicycle model (Gillespie, 1992).

The equations of motion with generalized coordinates as in (2.4) take the form:

$$\begin{aligned} m\left(\dot{V}'_x - \dot{\psi}V'_y\right) &= F'_{xf}c\delta - F'_{yf}s\delta + F'_{xr} - F'_w + G_x c\psi + G_y s\psi, \\ m\left(\dot{V}'_y + \dot{\psi}V'_x\right) &= F'_{xf}s\delta + F'_{yf}c\delta + F'_{yr} - G_x\, s\psi + G_y c\psi, \\ I_{z'}\ddot{\psi} &= \left(F'_{xf}\, s\delta + F'_{yf}c\delta\right) l_f - F'_{yr}l_r + M'_{sf} + M'_{sr}, \end{aligned} \tag{2.5}$$

where $I_{z'}$ – mass moment of inertia of the vehicle about the axis z', M'_{sf}, M'_{sr} – self-aligning moments, F'_{xf}, F'_{yf}, F'_{xr}, F'_{yr} – components of the tire-road interaction forces acting on the vehicle's wheels, F'_w – drag force, l_f, l_r – distance of the front and rear wheel axles from the vehicle's center of mass, G_x, G_y – components of the gravitational force (equal to zero when the road is not inclined).

To determine the tire-road interaction forces, formulas are used that relate the forces in the road plane to the friction coefficients and normal reactions.

Below are the verification results of both models through comparison of own calculations with results obtained using the CarSim software. The maneuver of a double lane change at a vehicle speed of 80 km/h was simulated (total simulation time $t_K = 12$ s, maximum lateral displacement of the vehicle at time $t = 4.9$ s was 3.69 m/s^2). Vehicle parameters were based on the CarSim A-Class Hatchback. The assumed front wheel steering angle is shown in Fig. 3a. Figure 3b presents the calculated vehicle trajectory.

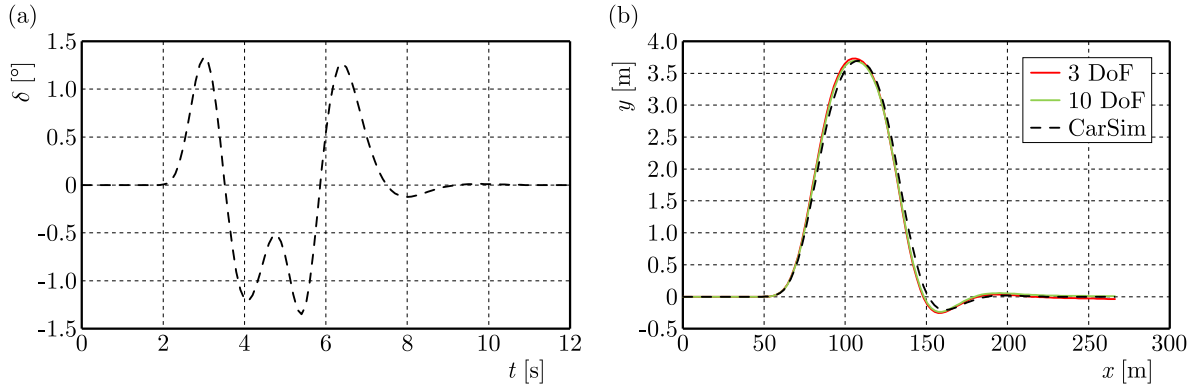

Fig. 3. Model validation: (a) assumed front wheel steering angle; (b) vehicle trajectory, own models and CarSim.

To assess the accuracy of the formulated models, the following error metrics were used:
– mean absolute error

$$\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n} |\varepsilon_i|\,, \tag{2.6}$$

– relative error

$$\varepsilon_{\%} = \left|\frac{w_o^{\max} - w_m^{\max}}{w_o^{\max}}\right| \cdot 100\,[\%], \tag{2.7}$$

where $\varepsilon_i = \varepsilon(t_i) = w_{o,i} - w_{m,i}$, $w_{o,i}$ – reference value obtained using the CarSim software, $w_{m,i}$ – value obtained according to the own model, $w_o^{\max} = \max\limits_{1\le i\le n} |w_{o,i}|$, $w_m^{\max} = \max\limits_{1\le i\le n} |w_{m,i}|$, n – number of compared values.

The results presented in Table 1 indicate that both own models yield errors $\bar{\varepsilon}$ in displacements on the order of several centimeters over a route nearly 270 meters long. The displacements and angular velocities ψ and $\dot{\psi}$ show larger errors between the own models and the CarSim software.

Table 1. Calculated mean $\overline{\varepsilon}$ and percentage $\varepsilon_{\%}$ differences between the own models and CarSim.

Model (DoF)	$\overline{\varepsilon}$				$\varepsilon_{\%}$			
	x [m]	y [m]	ψ [°]	$\dot{\psi}$ [°/s]	x	y	ψ	$\dot{\psi}$
3	0.150	0.057	0.219	0.392	0.06	0.97	2.65	2.05
10	0.150	0.052	0.217	0.381	0.06	0.04	2.11	0.90

However, these do not exceed 3 %. Therefore, both presented models can be considered valid. The 3 degrees of freedom model was previously verified in (Brzozowski & Drąg, 2023), and the 10 degrees of freedom model in (Brzozowski, 2025).

3. Path approximation

Proper path planning has a decisive impact on the path-following task (Zhong *et al.*, 2025; Guo *et al.*, 2025). There are many path planning methods. A review of these can be found in (Katrakazas *et al.*, 2015; Paden *et al.*, 2016). In this work, approximation using cubic B-spline functions (B3) was applied. It is assumed that the function approximating the path $f(x)$ has the form:

$$f(x) = \sum_{i=-1}^{n+1} a_i \varphi_i(x), \tag{3.1}$$

where a_i – coefficients, φ_i – cubic B-spline basis functions (B3), n – number of subintervals into which the interval $\langle A, B\rangle$ is divided. In the case where the interval $\langle A, B\rangle$ is divided into equal segments of length h, it takes the values:

$$x_i = ih \quad \text{for} \quad i = -3, -2, -1, 0, 1, ..., n, n+1, n+2, n+3, \tag{3.2}$$

functions $\varphi_i(x)$ are defined as follows:

$$\varphi_i(x) = \begin{cases} 0 & \text{when} \quad x < x_{i-2}, \\ (x - x_{i-2})^3 & \text{when} \quad x \in \langle x_{i-2}, x_{i-1}\rangle, \\ -3(x - x_{i-1})^3 + 3h(x - x_{i-1})^2 + 3h^2(x - x_{i-1}) + h^3 & \text{when} \quad x \in \langle x_{i-1}, x_i\rangle, \\ 3(x - x_{i+1})^3 + 3h(x - x_{i+1})^2 - 3h^2(x - x_{i+1}) + h^3 & \text{when} \quad x \in \langle x_i, x_{i+1}\rangle, \\ -(x - x_{i+2})^3 & \text{when} \quad x \in \langle x_{i+1}, x_{i+2}\rangle, \\ 0 & \text{when} \quad x > x_{i+2}. \end{cases} \tag{3.3}$$

The function $\varphi_i(x)$ along with its characteristic values is shown in Fig. 4.

The coefficients a_i present in Eq. (3.1) for $i = -1, 0, 1..., n, n+1$ are determined by minimizing the functional:

$$\Omega(a_{-1}, ..., a_{n+1}) = \sum_{k=0}^{m} [f(x_k) - y_k^e]^2 \to \min, \tag{3.4}$$

where y_k^e – measured value $y(x_k)$, $m+1$ – number of measurement points.

After transformations, to determine the $n+3$ coefficients $a_{-1}, ..., a_{n+1}$, a system of $n+3r$ linear algebraic equations of the form is obtained:

$$\sum_{i=-1}^{n+1} a_i \left[\sum_{k=0}^{m} \varphi_i(x_k^e)\, \varphi_j(x_k^e)\right] = \sum_{k=0}^{m} y_k^e \varphi_j(x_k^e) \tag{3.5}$$

for $j = -1, 0, 1, ..., n, n+1$.

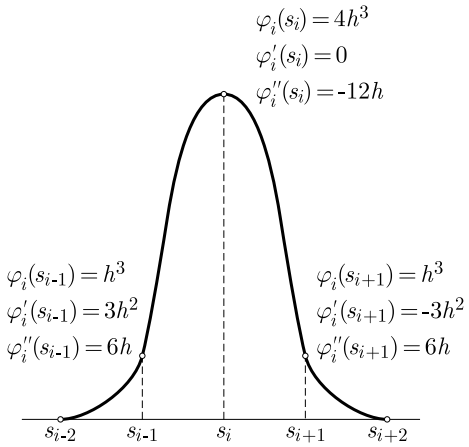

Fig. 4. Basis functions $\varphi_i(x)$.

This is a system of $n+3r$ linear algebraic equations. Solving this system enables the determination of the coefficients of the function f in formula (3.1). In problems related to approximating the vehicle's trajectory, the values of the function $f(x)$ and its derivatives at the beginning and end of the approximation interval are generally known, which enables the determination of up to 6 coefficients among $a_{-1}, ..., a_{n+1}$. To account for cases where the approximated path is not a function in the mathematical sense, the following procedure was applied.

If the points $(x_0^e, y_0^e), ..., (x_m^e, y_m^e)$ are sufficiently dense, the distance traveled by the vehicle can be approximately calculated as follows:

$$s_i^e = \sum_{j=1}^{i} \sqrt{\left(x_j^e - x_{j-1}^e\right)^2 + \left(y_j^e - y_{j-1}^e\right)^2}. \tag{3.6}$$

Assuming the vehicle speed is greater than zero, the values s_i^e form an increasing sequence. Therefore, the coordinates x and y can be treated as functions of the variable sss (in the mathematical sense). To determine the approximating functions $x(s)$ and $y(s)$, two problems analogous to the one presented above need to be solved, assuming:

$$\begin{aligned} x_i^e &= s_i^e \quad \text{and} \quad y_i^e = x_i^e \quad \text{when calculating the coefficients of the function } x(s), \\ x_i^e &= s_i^e \quad \text{and} \quad y_i^e = y_i^e \quad \text{when calculating the coefficients of the function } y(s). \end{aligned} \tag{3.7}$$

It is necessary to take into account the initial and boundary conditions for each of the functions $x(s)$, $y(s)$.

4. Own MPC/B algorithm for selecting the front wheel steering angle $\delta(t)$

We assume that the nonlinear equation describing the dynamics of the autonomous vehicle takes the form:

$$\mathbf{M}(\mathbf{q})\ddot{\mathbf{q}} = \mathbf{f}(\mathbf{q}, \dot{\mathbf{q}}, \delta(t)), \tag{4.1}$$

where $\mathbf{M}(\mathbf{q})$ – inertia matrix, $\mathbf{q}$, $\dot{\mathbf{q}}$ – vectors of vehicle coordinates and velocities, $\delta(t)$ – function describing the steering angle trajectory of the vehicle's front wheels (to be determined). It is assumed that the vehicle's path and velocity profile are known. The integration interval $\langle 0, T\rangle$ for the equations of motion (4.1) is divided into subintervals of length

$$\Delta t = mh_c, \tag{4.2}$$

where h_c – the integration step of the equations of motion, m_c – the multiple of the integration step.

Assuming that at $t = t_0$ the steering angle is known:

$$\delta_0 = \delta(t_0) \tag{4.3}$$

and the initial conditions:

$$q_0 = q(t_0), \qquad \dot{q}_0 = \dot{q}(t_0). \tag{4.4}$$

The sought value of the steering angle is denoted as

$$\delta_m = \delta(t_0 + \Delta t) = \delta(t_m) \tag{4.5}$$

which changes linearly over the interval $\langle t_0, t_0 + \Delta t\rangle = \langle t_0, t_m\rangle$ according to the relation:

$$\delta(t) = \delta_0 + \frac{\delta_m - \delta_0}{\Delta t}(t - t_0). \tag{4.6}$$

It is assumed that the quantity δ_m should minimize the expression:

$$\begin{aligned}\Delta^2(\delta_m) = {} & C_0^{x,y}\left[(x_{c,m} - x_{T,m})^2 + (y_{c,m} - y_{T,m})^2\right] + C_0^{\psi}[\psi_m - \psi_{T,m}]^2 \\ & + C_1^{x,y}\left[(\dot{x}_{c,m} - \dot{x}_{T,m})^2 + (\dot{y}_{c,m} - \dot{y}_{T,m})^2\right] + C_1^{\psi}[\dot{\psi}_m - \dot{\psi}_{T,m}]^2,\end{aligned} \tag{4.7}$$

where $x_{c,m} = x_c(t_m)$, $y_{c,m} = y_c(t_m)$, $\psi_m = \psi(t_m)$. They are the coordinates of the vehicle's center of mass and its yaw angle, calculated using the vehicle dynamics model at time t_m, whereas $x_{T,m} = x_T(t_m)$, $y_{T,m} = y_T(t_m)$, $\psi_{T,m} = \psi_T(t_m)$. They are the desired path coordinates and the tangent angle to the vehicle trajectory, calculated according to the path approximation algorithm at time t_m.

To calculate $\Delta_m^2(\delta_m)$, the equation of motion (4.1) must be integrated with $\delta(t)$ defined by (4.6). The quadratic form (4.7) reaches its minimum when:

$$\begin{aligned}\frac{\partial \Delta^2(\delta_m)}{\partial \delta} = 2\Bigg\{ & C_0^{x,y}\left[(x_{c,m} - x_{T,m})\frac{\partial x_{c,m}}{\partial \delta} + (y_{c,m} - y_{T,m})\frac{\partial y_{c,m}}{\partial \delta}\right] + C_0^{\psi}(\psi_m - \psi_{T,m})\frac{\partial \psi_m}{\partial \delta} \\ & + C_1^{x,y}\left[(\dot{x}_{c,m} - \dot{x}_{T,m})\frac{\partial \dot{x}_{c,m}}{\partial \delta} + (\dot{y}_{c,m} - \dot{y}_{T,m})\frac{\partial \dot{y}_{c,m}}{\partial \delta}\right] \\ & + C_1^{\psi}\left(\dot{\psi}_m - \dot{\psi}_{T,m}\right)\frac{\partial \dot{\psi}_m}{\partial \delta}\Bigg\} = 0,\end{aligned} \tag{4.8}$$

where $\frac{\partial p}{\partial \delta}\, p \in \Omega = \left\{x_{c,m}, y_{c,m}, \psi_m, \dot{x}_{c,m}, \dot{y}_{c,m}, \dot{\psi}_m\right\}$ is the derivative of the function p with respect to δ for $t = t_m$.

To calculate these quantities in this paper, the five-point finite difference method was applied, assuming:

$$\frac{\partial p}{\partial \delta}_{\delta=\delta_m} = \frac{p(\delta_0 - 2\delta_m) - 8p(\delta_0 - \delta_m) + 8p(\delta_0 + \delta_m) - p(\delta_0 + 2\delta_m)}{12\delta_m}. \tag{4.9}$$

To calculate $\frac{\partial p}{\partial \delta}$ for $p \in \Omega t$, it is therefore necessary to quadratically integrate the equation of motion $(3.7)_1$ in the interval $\langle t_0, t_m\rangle$.

The δ_m is determined using Newton's successive approximation method, assuming:

$$\delta_m^0 = \delta_0, \qquad \delta_m^{(i)} = \delta_m^{(i-1)} - \frac{\gamma_i(\delta_m)}{\gamma_i'(\delta_m)}, \tag{4.10}$$

where

$$\gamma_i(\delta_m) = \frac{1}{2}\frac{\partial \Delta^2(\delta_m)}{\partial \delta}, \qquad \gamma_i'(\delta_m) = \frac{\partial \gamma_i(\delta_m)}{\partial \delta}.$$

The iterative process was conducted until one of the conditions was met:

$$i = i_{\mathrm{MAX}}, \qquad \left|\frac{\gamma_i(\delta_m)}{\gamma_i'(\delta_m)}\right| < \mathrm{EPS}, \tag{4.11}$$

where i_{MAX} and EPS are quantities defining the maximum number of iterations and the absolute error in determining the δ_m, respectively.

The derivative of $\gamma_i'(\delta_m)$ was also calculated using the five-point finite difference method.

In summary, to determine δ_m, it is necessary, according to the proposed algorithm, to integrate the equations of motion (4.1) over the interval $\langle t_0, t_m \rangle$ at most:

$$N = i_{\mathrm{MAX}} 5\mathrm{times}. \tag{4.12}$$

5. Simulation research

Simulation studies were performed for a loop in which the trajectory (Fig. 5) is described by the formulae:

$$\begin{cases} x = a\sin(s), \\ y = a\sin(s)\cos(s), \end{cases}$$

where $a = 50$.

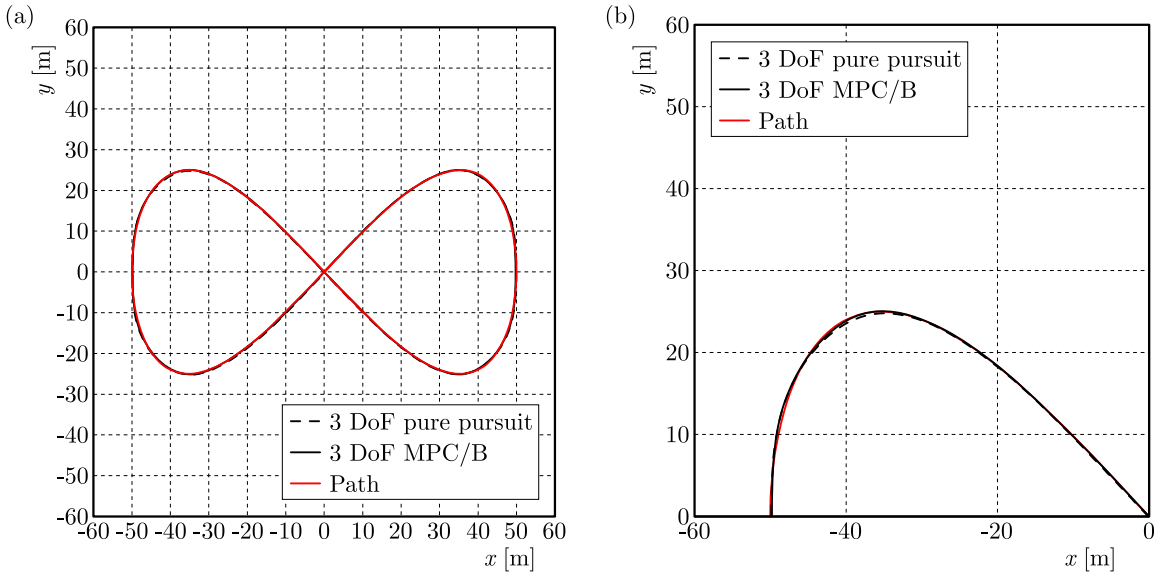

Fig. 5. Trajectory for the implementation of the "loop" maneuver: (a) total; (b) enlargement.

The execution time for the entire maneuver is $t_k = 39\,\mathrm{s}$, and the total distance is 304 metres. A constant speed of $v = 28\,\mathrm{km/h}$ was assumed.

In the present task, the constants $C_0^{x,y}$ and $C_1^{x,y}$ are taken as 10^3 and 10^2, while the constant L_d needed for the PP algorithm to work properly as 0.05.

The trajectory shown in Fig. 5 indicates that there is little difference between control using the PP algorithm and the proprietary algorithm. The proposed algorithm is slightly more accurate. Due to the small differences, the figure does not show the results of the calculation using the dynamics model with 10 DoF. The maximum values of the mapping error are shown in Table 2.

Table 2. Mapping error.

Dynamics model	Steering algorithm	
	PP	MPC/B
3 DoF	0.298	0.098
10 DoF	0.409	0.384

The results indicate a higher accuracy of the proposed algorithm, especially when combined with a low complexity vehicle dynamics model. Figure 6 shows the course of the steering angle. The MPC/B algorithm determined a slightly larger steering angle than the PP algorithm. The maximum difference was 1.64°.

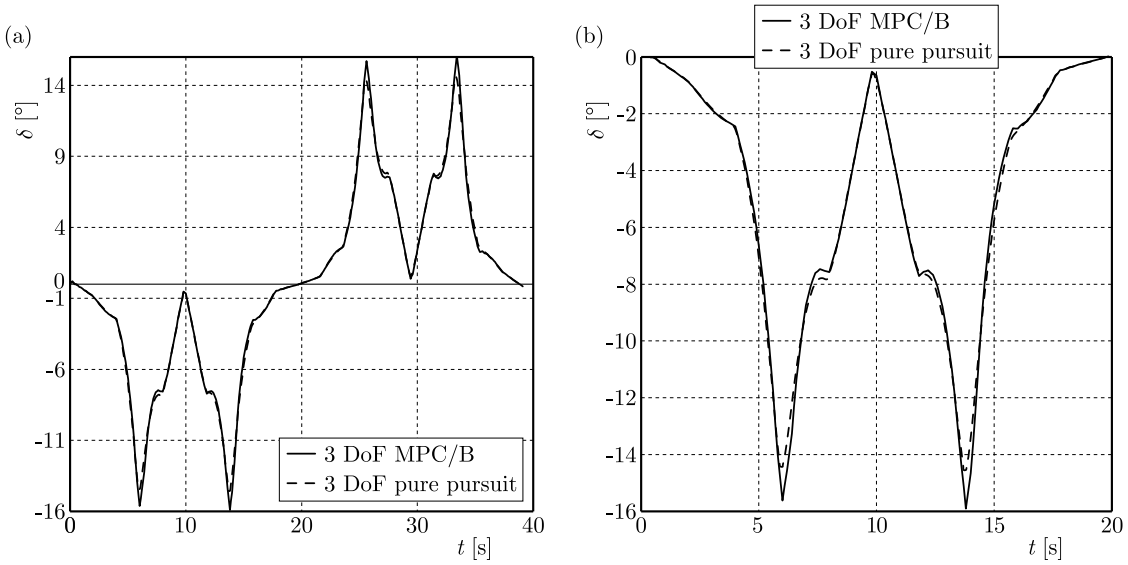

Fig. 6. Diagram of the course of the steering angle: (a) total; (b) magnification.

6. Results and discussion

Among the most popular vehicle control methods are: geometric, model-based (optimization) and machine learning. Developing control methods that ensure path realization is an important research problem. This paper compares the classical geometric algorithm PP with the proprietary MPC/B algorithm, which belongs to the group of MPC algorithms based on a model of vehicle dynamics. The proposed algorithm is more accurate than the PP algorithm. It has additional advantages, such as the ability to be used with any dynamics model or to be tuned for specific requirements. A weakness is the moderate computational efficiency. For the presented "loop" maneuver, the computation time with the PP algorithm was 2.14 s for MPC/B 4.94 s. Table 3 shows a synthesis of the conclusions and a comparison of the PP algorithm with the proposed own algorithm.

Table 3. Comparison of the PP algorithm and own algorithm MPC/B.

Dynamics model	PP	MPC/B
	Does not use	Any of the following may be used
Choice of constants	Limited tuning ability	Trajectory or yaw angle tuning possible
Precision	Moderate	High
Numerical effectiveness	High	Moderate

In future work, it seems expedient to compare the proposed algorithm with a standard MPC-type algorithm.

References

1. Ajanović, Z., Regolin, E., Shyrokau, B., Ćatić, H., Horn, M., & Ferrara, A. (2023). Search-based task and motion planning for hybrid systems: Agile autonomous vehicles. *Engineering Applications of Artificial Intelligence*, *121*, Article 105893. https://doi.org/10.1016/j.engappai.2023.105893

2. Amer, N.H., Zamzuri, H., Hudha, K., Aparow, V.R., Kadir, Z.A., & Abdin, A.F.Z. (2018). Path tracking controller of an autonomous armoured vehicle using modified Stanley controller optimized with particle swarm optimization. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, *40*(2), Article 104. https://doi.org/10.1007/s40430-017-0945-z

3. An, L., Huang, X., Yang, P., & Liu, Z. (2025). Adaptive Bézier curve-based path following control for autonomous driving robots. *Robotics and Autonomous Systems*, *189*, Article 104969. https://doi.org/10.1016/j.robot.2025.104969

4. Blajer, W. (1998). *Methods of multibody system dynamic* (in Polish). Radom: Politechnika Radomska.

5. Brzozowski, M. (2025). *Motion planning for autonomous vehicles using dynamic models* (in Polish) [Unpublished doctoral dissertation]. Faculty of Mechanical Engineering and Computer Science, University of Bielsko-Biala, Poland.

6. Brzozowski, M., & Drąg, Ł. (2023). Application of dynamic optimization for autonomous vehicle motion control. *Transport Problems*, *18*(2), 209-222. https://doi.org/10.20858/tp.2023.18.2.18

7. Cao, Y., Ni, K., Jiang, X., Kuroiwa, T., Zhang, H., Kawaguchi, T., Hashimoto, S., & Jiang, W. (2023). Path following for autonomous ground vehicle using DDPG algorithm: A reinforcement learning approach. *Applied Sciences*, *13*(11), Article 6847. https://doi.org/10.3390/app13116847

8. Cibooglu, M., Karapinar, U., & Söylemez, M.T. (2017). Hybrid controller approach for an autonomous ground vehicle path tracking problem. *2017 25th Mediterranean Conference on Control and Automation (MED)* (pp. 583–588). IEEE. https://doi.org/10.1109/MED.2017.7984180

9. Coulter, R.C. (1992). *Implementation of the pure pursuit path tracking algorithm* (Technical Report CMU-RI-TR-92-01). Robotics Institute, Carnegie Mellon University. https://www.ri.cmu.edu/pub_files/pub3/coulter_r_craig_1992_1/coulter_r_craig_1992_1.pdf

10. Diachuk, M., & Easa, S.M. (2022). Motion planning for autonomous vehicles based on sequential optimization. *Vehicles*, *4*(2), 344–374. https://doi.org/10.3390/vehicles4020021

11. Elsayed, H., Abdullah, B.A., & Aly, G. (2018). Fuzzy logic based collision avoidance system for autonomous navigation vehicle. *2018 13th International Conference on Computer Engineering and Systems (ICCES)* (pp. 469–474). IEEE. https://doi.org/10.1109/ICCES.2018.8639396

12. Fu, T., Zhou, H., & Liu, Z. (2022). NMPC-based path tracking control strategy for autonomous vehicles with stable limit handling. *IEEE Transactions on Vehicular Technology*, *71*(12), 12499–12510. https://doi.org/10.1109/TVT.2022.3196315

13. Gámez Serna, C., Lombard, A., Ruichek, Y., & Abbas-Turki, A. (2017). GPS-based curve estimation for an adaptive pure pursuit algorithm. In G. Sidorov, & O. Herrera-Alcántara (Eds.), *Lecture notes in computer science: Vol. 10061. Advances in computational intelligence* (pp. 497–511). Springer. https://doi.org/10.1007/978-3-319-62434-1_40

14. Gillespie, T.D. (1992). *Fundamentals of vehicle dynamics*. SAE International.

15. Guo, S., Gong, J., Shen, H., Yuan, L., Wei, W., & Long, Y. (2025). DBVSB-P-RRT*: A path planning algorithm for mobile robot with high environmental adaptability and ultra-high speed planning. *Expert Systems with Applications*, *266*, Article 126123. https://doi.org/10.1016/j.eswa.2024.126123

16. Huang, P., Zhang, Z., Luo, X., Zhang, J., & Huang, P. (2018). Path tracking control of a differential-drive tracked robot based on look-ahead distance. *IFAC-PapersOnLine*, *51*(17), 112–117. https://doi.org/10.1016/j.ifacol.2018.08.072

17. Katrakazas, C., Quddus, M., Chen, W-H., & Deka, L. (2015). Real-time motion planning methods for autonomous on-road driving: State-of-the-art and future research directions. *Transportation Research Part C: Emerging Technologies*, *60*, 416–442. https://doi.org/10.1016/j.trc.2015.09.011

18. Lee, K., Jeon, S., Kim, H., & Kum, D. (2019). Optimal path tracking control of autonomous vehicle: Adaptive full-state linear quadratic Gaussian (LQG) control. *IEEE Access*, *7*, 109120–109133. https://doi.org/10.1109/ACCESS.2019.2933895

19. Li, S., Wang, G., Zhang, B., Yu, Z., & Cui, G. (2019). Vehicle stability control based on model predictive control considering the changing trend of tire force over the prediction horizon. *IEEE Access*, *7*, 6877–6888. https://doi.org/10.1109/ACCESS.2018.2889997

20. Liu, J., Yang, Z., Huang, Z., Li, W., Dang, S., & Li, H. (2021). Simulation performance evaluation of pure pursuit, Stanley, LQR, MPC controller for autonomous vehicles. *2021 IEEE International Conference on Real-time Computing and Robotics (RCAR)* (pp. 1444–1449). IEEE. https://doi.org/10.1109/RCAR52367.2021.9517448

21. Pacejka, H.B., & Sharp, R.S. (1991). Shear force development by pneumatic tyres in steady state conditions: A review of modelling aspects. *Vehicle System Dynamics*, *20*(3–4), 121–175. https://doi.org/10.1080/00423119108968983

22. Paden, B., Čáp, M., Yong, S.Z., Yershov, D., & Frazzoli, E. (2016). A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Transactions on Intelligent Vehicles*, *1*(1), 33–55. https://doi.org/10.1109/TIV.2016.2578706

23. Rajamani, R. (2012). *Vehicle dynamics and control*. Springer. https://doi.org/10.1007/978-1-4614-1433-9

24. Yang, J., Bao, H., Ma, N., & Xuan, Z. (2017). An algorithm of curved path tracking with prediction model for autonomous vehicle. *2017 13th International Conference on Computational Intelligence and Security (CIS)* (pp. 405–408). IEEE. https://doi.org/10.1109/CIS.2017.00094

25. Yang, Y., Li, Y., Wen, X., Zhang, G., Ma, Q., Cheng, S., Qi, J., Xu, L., & Chen, L. (2022). An optimal goal point determination algorithm for automatic navigation of agricultural machinery: Improving the tracking accuracy of the Pure Pursuit algorithm. *Computers and Electronics in Agriculture*, *194*, Article 106760. https://doi.org/10.1016/j.compag.2022.106760

26. Zhang, B., Zong, C., Chen, G., & Li, G. (2019). An adaptive-prediction-horizon model prediction control for path tracking in a four-wheel independent control electric vehicle. *Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering*, *233*(12), 3246–3262. https://doi.org/10.1177/0954407018821527

27. Zhong, J., Kong, D., Wei, Y., Hu, X., & Yang, Y. (2025). Efficiency-optimized path planning algorithm for car-like mobile robots in bilateral constraint corridor environments. *Robotics and Autonomous Systems*, *186*, Article 104923. https://doi.org/10.1016/j.robot.2025.104923

Manuscript received June 3, 2025; accepted for publication July 10, 2025; published online September 2, 2025.

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 781–794, Warsaw 2025
https://doi.org/10.15632/jtam-pl/208484

IMPACT OF HONEYCOMB STRAIGHTENER PARAMETERS ON OPERATION IN A STRAIGHT DUCT

Emil SMYK[1*], Michał STOPEL[1], Adam RACHWALSKI[2]

[1] *Bydgoszcz University of Science and Technology, Bydgoszcz, Poland*
[2] *Hanplast Sp. z o.o., Bydgoszcz, Poland*

*corresponding author, *emil.smyk@pbs.edu.pl*

The influence of the honeycomb diameter and straightener length on performance was investigated. Velocity profiles were measured using a hot-wire anemometer, and pressure losses were also recorded. The straighteners were placed 10D downstream of the fan. Measurements were conducted at Reynolds numbers of 10000, 15000, 30000, 45000. Additionally, two methods were proposed to assess the influence of straighteners on the shape of the velocity profile. The results showed that at Reynolds numbers of 10000 and 15000, straighteners had only a minor effect on reducing turbulence intensity and relaminarizing the velocity profile. In contrast, at higher Reynolds numbers, their impact was significant.

Keywords: pressure drop; Fanning friction factor; head pressure losses; channels; inner flow.

1. Introduction

Stream straighteners are simple devices used in ventilation ducts, HVAC systems, jet engines, industrial pipelines, scientific research, tanks, and overflow vessels. They are designed to correct unexpected and unwanted flow profile effects, often caused by the absence of required straight sections upstream and downstream of the measurement point. Also known as stream conditioners, these devices reduce hydraulic entrance length and decrease turbulence intensity, enabling the required velocity profile. Some types of stream straighteners have been described in (International Organization for Standardization, 2022). However, due to their simple manufacturing process, the most common flow straighteners are those in the shape of a honeycomb (Hrúz *et al.*, 2020).

The earliest paper on flow straighteners found is the one by Bradshaw (1965). He cited earlier work (from 1959), but access to it could not be obtained. He investigated wind tunnel screens. Lumley and McMahon (1967) demonstrated that this straightener significantly reduced turbulence intensity. Tan-Atichat *et al.* (1982) proposed using screens and grids to decrease turbulence, showing that an appropriately selected screen and a grid (acting as a turbulence generator) could shorten turbulence decay distance by a factor of four. Groth and Johansson (1988) examined a cascade of the screens in a wind tunnel and concluded that using a major loss factor as an efficiency determinant was incorrect since different screen arrangements with

Ministry of Science and Higher Education
Republic of Poland

This publication has been funded by the Polish Ministry of Science and Higher Education under the Excellent Science II programme "Support for scientific conferences".
The content of this article was presented during the 61st Symposium "Modelowanie w mechanice" (Modelling in Mechanics), Szczyrk, Poland, March 2–5, 2025.

varying major loss factors could yield the same turbulence reduction. Laws (1990) introduced a new type of straightener/conditioner featuring holes of different diameters in the outer and inner rings. He suggested that the minimum straightener length should be D/8, where D is the channel diameter, and noted that straightener length does not significantly affect performance.

Xiong *et al.* (2003) compared two perforated plates and a tube bundle, showing that perforated plates act as turbulence grids, producing homogeneous and quasi-isotropic turbulence more efficiently than tube bundles. The turbulence field does not reach equilibrium even at a downstream position of 50 diameters. Saunders *et al.* (2004) tested honeycomb straighteners with screens in a wind tunnel and, like Xiong *et al.* (2003), found that turbulence initially increased downstream of the flow conditioners before decreasing. Hamzah *et al.* (2021) numerically investigated a honeycomb straightener in a wind tunnel and found that straighteners improved flow parameters, increasing the usable test. However, they simulated the straightener as a porous medium, making their results applicable only to simulation.

Kühnen *et al.* (2018) demonstrated that flow relaminarization can be achieved by reducing wall shear using two specially designed passive conditioners. El Drainy *et al.* (2009) studied tangential vortex induction behind the Zenker plates, showing that tangential velocity depends on plate thickness and disappears with increased straightener length. Sun *et al.* (2023; 2025) investigated the role of the honeycomb in the relaminarization of the synthetic jet. They show that the honeycomb straightener can reduce the periodic and random velocity fluctuation even in periodic phenomena such as synthetic jets. They also pointed out the need to properly select the length and diameter of the straightener. Jurga *et al.* (2024) investigated honeycomb straighteners both upstream and downstream of an elbow. Their studies demonstrated that the straightener suppresses the secondary flow generated by the elbow and positively influences the velocity profile. Flow through an elbow is one of the fundamental types of configuration analyzed in the literature, for example, in (Dutta *et al.*, 2025; Smyk *et al.*, 2024).

Kühnen *et al.* (2018) suggest that the objective of flow relaminarization is to minimize energy consumption, given that laminar flow results in reduced losses compared to turbulent flow. Apart from reduced flow losses, a symptom of laminar flow is the shape of the velocity profile, expressed by Prandtl's power law formula (Salama, 2021). Kühnen *et al.* (2018) also suggest that a paraboloidal velocity profile indicates relaminarization of the flow. Flow relaminarization is achieved by locally increasing shear stresses (Jurga *et al.*, 2024; Kühnen *et al.*, 2018), and can be classified as a passive flow control method. Flow straighteners are similar in design to filters and consist of small channels (Kaminski *et al.*, 2025). These can also be used as stream straighteners. Passive flow control methods involving the local increase of shear stresses are also employed in external flows (Dróżdż *et al.*, 2025; Klotz *et al.*, 2024), where the objective is to attain an optimal velocity profile shape.

In the analyzed papers, the straighteners were investigated mainly in wind tunnels and the turbulence and profile disturbance were generated very often by the grid. Flow straighteners are mainly used in pipelines, and based on the literature, it is difficult to indicate an unambiguous way of designing and the scope of application of straighteners. For example, to determine what Reynolds numbers of flow they should be used for. The purpose of this article is to discuss the influence of the channel diameter and length of honeycomb straighteners on their performance. The study also used a different vortex generator than in the remaining literature, as it was an axial fan, which is more consistent with the structure and characteristics of real-life cases of using straighteners in industry. The data were presented in a manner that makes it possible to create a numerical model and extend the research to additional cases.

2. Materials and methods

The impact of the honeycomb straighteners on flow parameters was assessed using a test channel equipped with measurement devices, as shown in Fig. 1. Honeycombs were manufac-

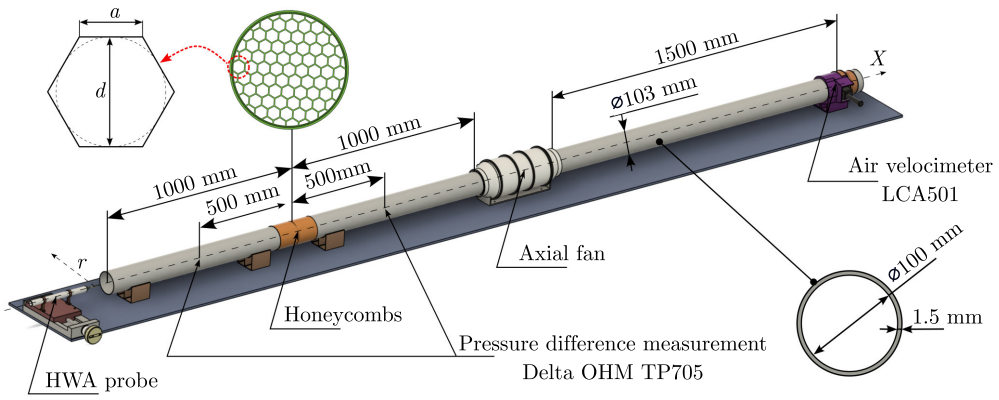

Fig. 1. Test channel schematic with the measurement equipment.

tured with 3D print technology with polylactide acid (PLA) and placed in the straight channel 1000 mm behind the axial fan (Soler&Palau TD-250) controlled by a Soler&Palau REB-1 speed controller. The fan acted as a turbulence generator. To evaluate the effect of straighteners on the flow parameters, the velocity profile in the channel was measured using a MiniCTA 55T30 hot-wire anemometer (HWA) with a 55P16 single-wire probe and a NI9215 data acquisition device. Velocity profiles were measured 1000 mm downstream of the honeycomb. The anemometer was calibrated for velocity measurements ranging from 1.5 m/s to 26 m/s. Temperature correction was applied, and the velocity measurement accuracy was within ±6 % of the reading. The HWA probe was placed 1 mm behind the channel outlet, and positioning was performed using a micrometer screw. The inner diameter of the pipe (channel diameter) was $D = 100$ mm, and the outer diameter of the pipe was $D_{\text{out}} = 103$ mm. The pressure drop on the honeycomb was measured with the differential pressure probe Delta OHM TP705-10MBD (the measurement range was 1000 Pa, the accuracy was ±0.5 % of full scale, and the measurement resolution was 1 Pa) connected to Delta OHM DH 2124.2 pressure meter. The pressure drop was measured at a distance of 1000 mm, as shown in Fig. 1. Mean velocity was determined using an Airflow LCA501 air velocity meter (measurement range: 0.25 m/s–30 m/s, accuracy: ±1 % of reading) placed 1500 mm upstream of the fan.

Measurements were conducted for six honeycomb straighteners and a control channel without straighteners. The parameters of the honeycomb straighteners are presented in Table 1. They covered four different Reynolds numbers Re = 10000, 15000, 30000, 45000, calculated as

$$\text{Re} = \frac{UD}{\nu}, \tag{2.1}$$

where U is a mean velocity [m/s], and ν is a kinematic viscosity equal of $\nu = 15.16 \cdot 10^{-6}$ m^2/s for the air temperature of 21 °C and the air humidity of 37 %.

Table 1. Dimensions of the honeycomb straighteners.

Case	Honeycomb diameter d [mm]	Honeycomb side length a [mm]	Straightener length L [mm]
Without	–	–	–
D10L5	10	5.77	5
D10L10	10	5.77	10
D10L20	10	5.77	20
D10L40	10	5.77	40
D5L20	5	2.89	20
D20L20	20	11.55	20

Velocity profiles were measured directly at the fan outlet and at a distance of 1000 mm behind the fan (where the straighteners were installed to illustrate canes in flow parameters within a straight channel).

The velocity measurement with the HWA was conducted at a sampling frequency of $f = 50\,\text{kHz}$ for 1 s at each measurement point. This setup allowed both velocity and turbulence intensity measurements. The mean velocity (U), standard deviation of the velocity (U_{rms}) and the turbulence intensity (Tu) of the flow were determined as

$$U = \frac{1}{N}\sum_{i=1}^{N} u_i, \qquad U_{\text{rms}} = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(u_i - U)^2}, \qquad Tu = \frac{U_{\text{rms}}}{U}, \tag{2.2}$$

where u_i is a measured velocity series sample [m/s], N is the number of series samples ($N = 50000$).

This study investigated the impact of honeycomb straighteners on the velocity profile. To analyze this effect, Prandtl's power law formula was applied, which describes the velocity profile in turbulent flow (Salama, 2021):

$$\frac{U}{U_{\max}} = \left(1 - \frac{r}{R}\right)^{1/n}, \tag{2.3}$$

where $U_{\max}$ is the maximum velocity value [m/s], r is the radial distance from the axis (see Fig. 1) [m], R is a channel radius ($R = D/2 = 50\,\text{mm}$), n is an exponent depending on the Reynolds number. The exponent n was calculated for each measured velocity profile in such a way as to satisfy the relationship:

$$\left(\frac{1}{U_{\max}^2}\sum\left(U(r)_{\text{experimental}} - U(r)_{\text{theoretical}}\right)^2\right) \to 0, \tag{2.4}$$

where $U_{\text{theoretical}}$ was calculated from Eq. (2.3). The Solver add-in in Microsoft Excel was used to find the value of exponent n.

3. Results

3.1. Profiles in a straight channel

Figure 2 presents the velocity profile measured directly on the outlet of the fan (Fig. 2a), 1000 mm downstream of the fan (Fig. 2b), and 2000 mm downstream of the fan (Fig. 2c). The change in the velocity profile shape and the turbulence intensity with increasing distance from the fan was evident for all Reynolds numbers. The velocity profile becomes more rectangular as the distance from the fan increases. Additionally, for Re = 10000 a distinct rounding of the profile near the duct wall is observed. Generally, higher Reynolds numbers correspond to flatter velocity profiles (Salama, 2021). However, as shown in Fig. 2c, the velocity profiles for Re = 10000 and 15000 appear flatter than those for Re = 30000 and 45000. This discrepancy is attributed to greater irregularities in the velocity profile at high Reynolds numbers compared to lower ones. It is expected that using a longer channel would result in further profile flattering at high Reynolds numbers.

As expected, higher Reynolds numbers correspond to increased turbulence levels. The highest turbulence intensity was observed near the walls and was due to shear effects (Hwang, 2024). The turbulence intensity near the wall, relative to the duct axis, is evident for Re = 10000, 15000, and 30000. However, for all cases, turbulence intensity decreased with increasing distance from the fan. The change in turbulence intensity at the duct axis between the fan outlet and 2000 mm downstream is as follows: from 2.78 % to 0.86 % for Re = 10000; from 3.68 % to 0.52 % for Re = 15000; from 17.84 % to 5.85 % for Re = 30000; from 13.52 % to 5.93 % for Re = 45000.

The direct correlation between turbulence intensity and the Reynolds number confirms the feasibility of using a duct fan as a vortex generator in the flow.

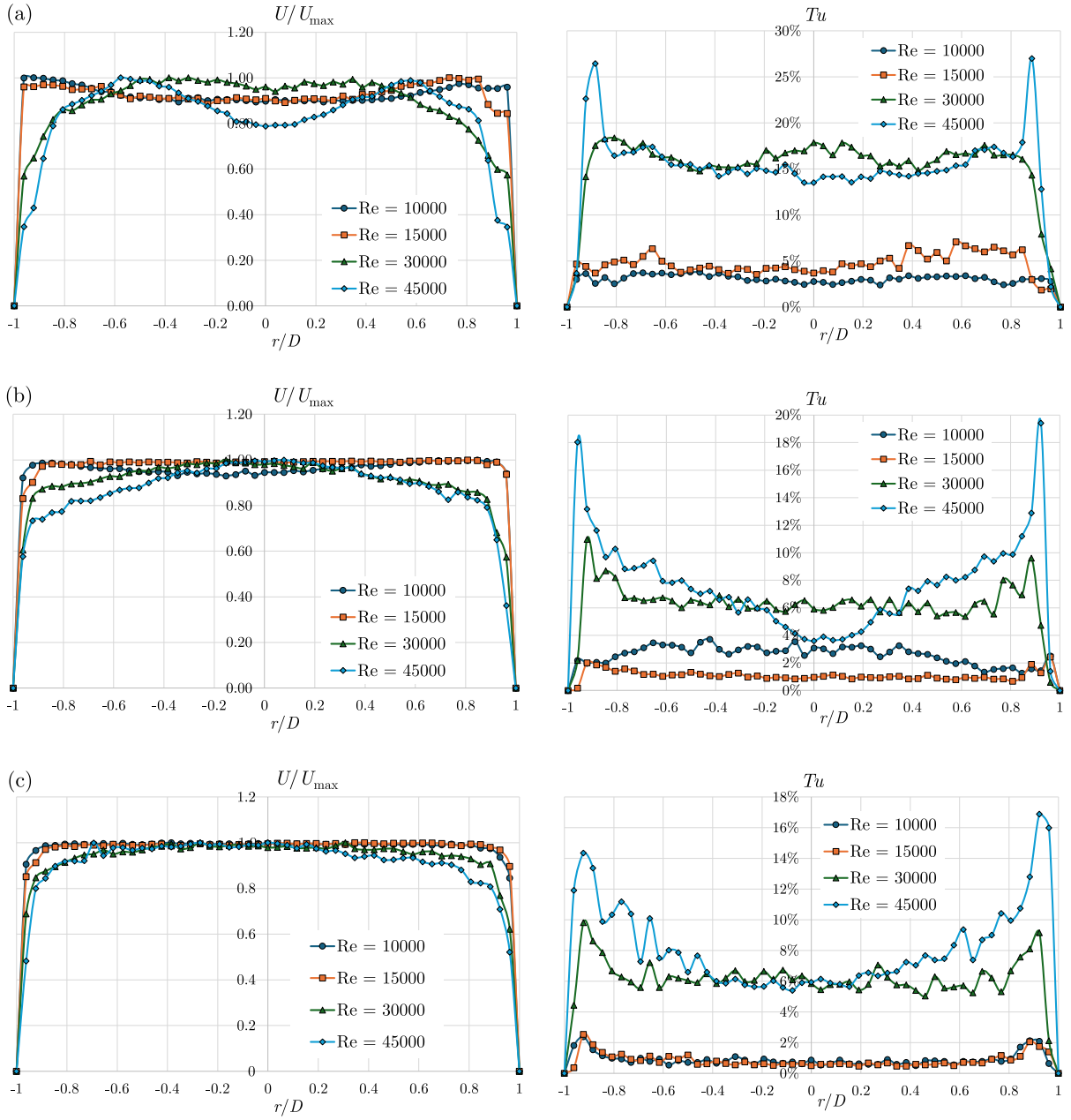

Fig. 2. Velocity (left) and turbulence intensity (right) in the outlet of the fun (a), 1000 mm behind the fun (b), and 2000 mm behind the fun (c).

3.2. Profiles at different straightener lengths

Figure 3 presents the velocity and turbulence profile measured 1000 mm downstream of a straightener for different straightener lengths and Reynolds numbers. At low Reynolds numbers (Figs. 3a and 3b), the impact of the straightener is minimal, with velocity profiles remaining nearly identical regardless of the straightener's length or presence. In terms of turbulence intensity, minor yet noticeable changes were observed. A significant reduction in turbulence intensity was detected near the duct wall at Re = 10000. At this Reynolds number, turbulence intensity decreased for all tested straightener lengths. However, the change was not substantial, as turbulence intensity remained below 1 % even without a straightener. At Re = 15000, no turbulence intensity was observed near the wall. However, for straighteners of 10 mm or longer, a decrease of approximately 0.25 percentage points in turbulence intensity was noted near the axis ($-0.4 < r/D < 0.4$).

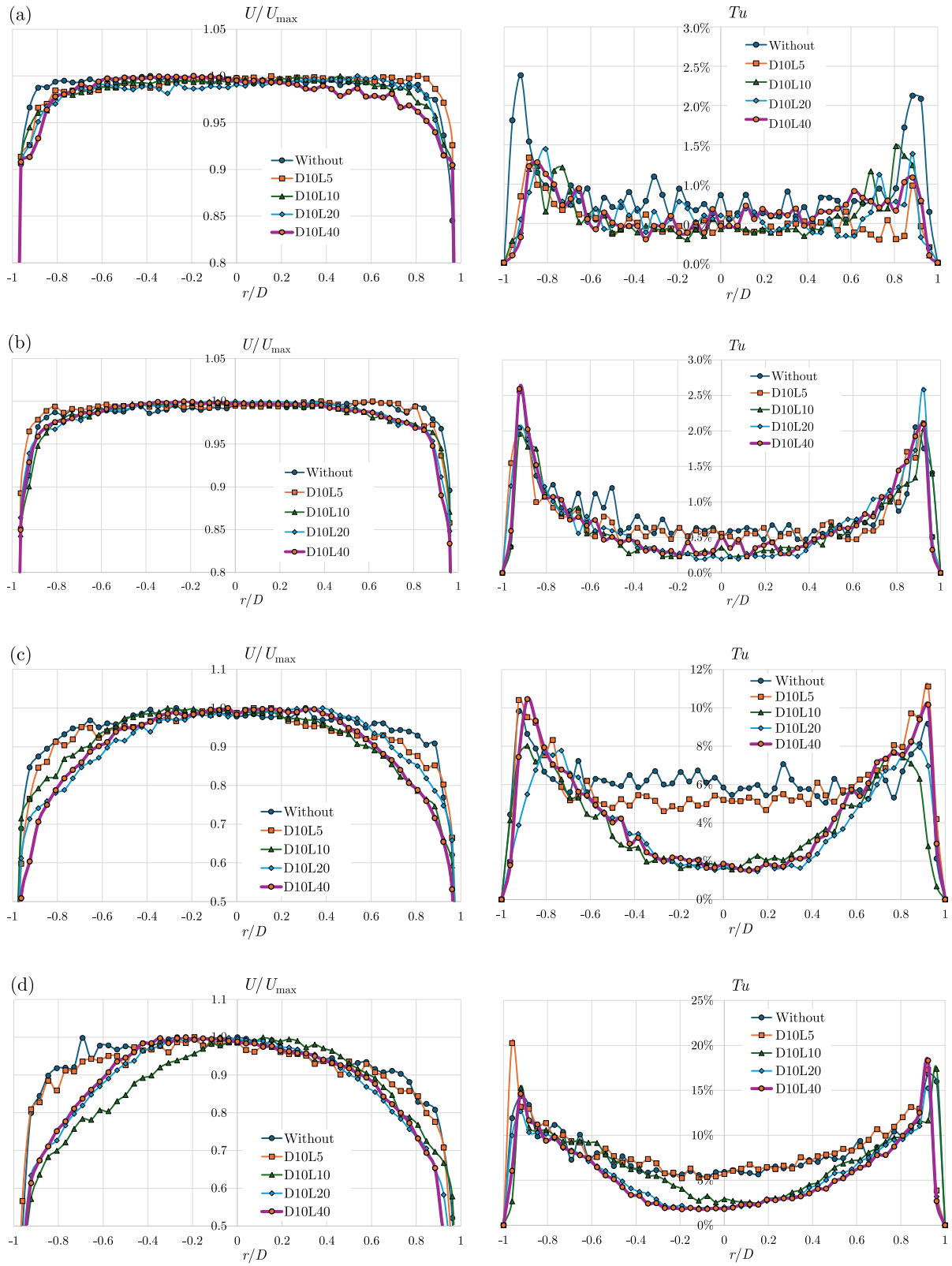

Fig. 3. Velocity (left) and turbulence intensity (right) profile at different Reynolds numbers and straightener lengths: (a) Re = 10000; (b) Re = 15000; (c) Re = 30000; (d) Re = 45000.

The use of the honeycomb straighteners had a significant influence on the shape of the velocity profile at Re = 30000. The longer the straightener, the more rounded and less squared the measured velocity profile becomes. This is evident from the reduced velocity values near

the duct walls ($r/D > \pm 0.4$). Turbulence intensity near the axis decreased for all investigated straighteners – by about 1 percentage point for L10D5, and approximately 4 percentage points for the remaining configurations. However, an increase in turbulence intensity near the duct walls was observed for D10L5 and D10L40.

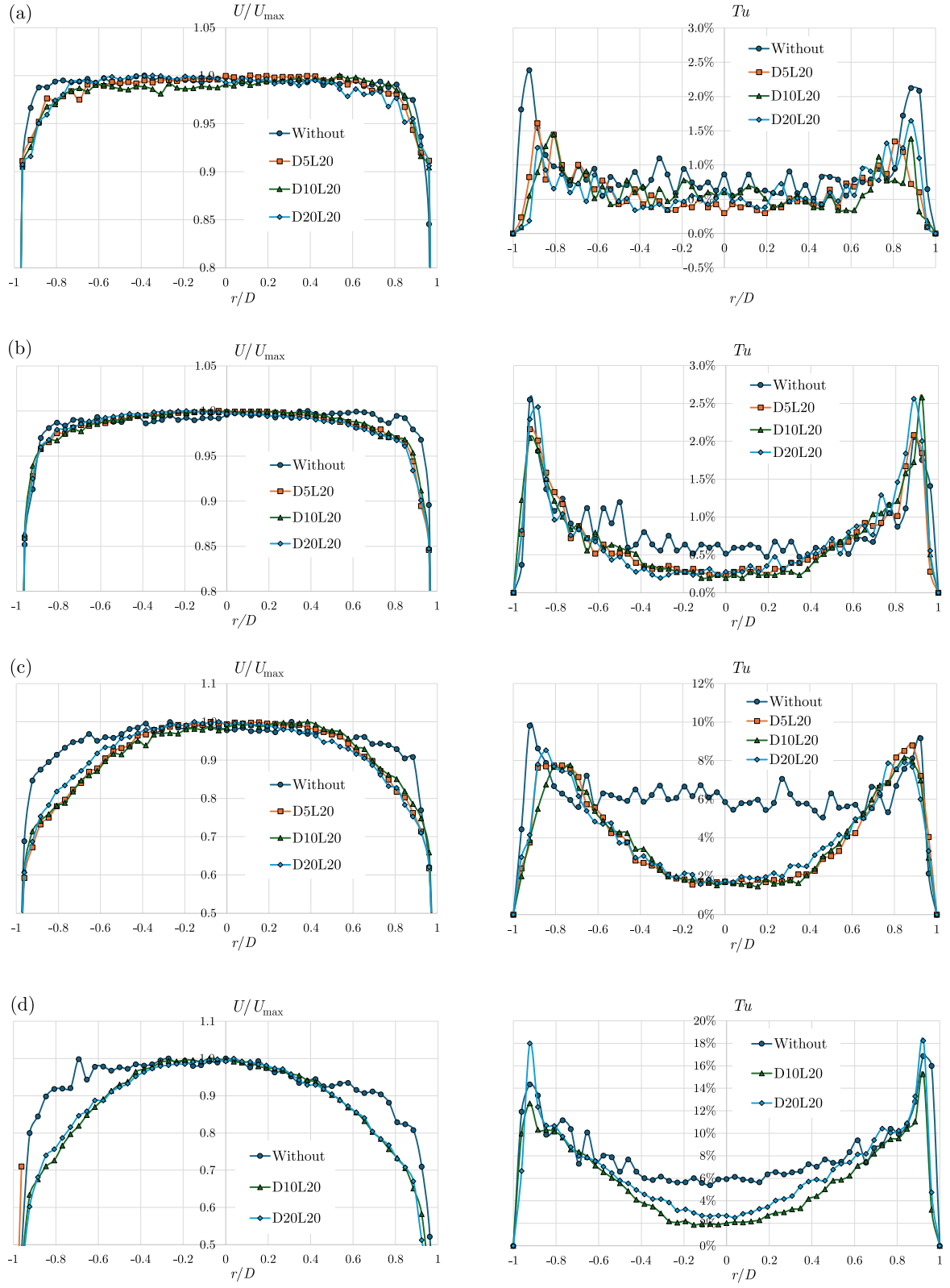

Fig. 4. Velocity (left) and turbulence intensity (right) profile at different Reynolds numbers and honeycomb diameters: (a) Re = 10000; (b) Re = 15000; (c) Re = 30000; (d) Re = 45000.

At Re = 45000, the impact of honeycomb straighteners on the velocity profiles was similar to that at Re = 30000. Again, longer straighteners resulted in more rounded velocity profiles, while the L10D5 configuration produced a noticeably asymmetric profile. Turbulence intensity decreased for all cases except the shortest straightener D10L5. Along the axis, the reduction in turbulence intensity decreased by about 3.03 percent point in the case of L10D10, 3.91 percent point in the case of L10D20, and 4.10 percent point in the case of L10D40. The profile asymmetry observed in the D10L10 case was reflected in its turbulence intensity.

3.3. Profiles at different honeycomb diameters

The velocity and turbulence intensity profiles at different Reynolds numbers and honeycomb sizes are presented in Fig. 4. At Re = 10000, the straighteners had no significant impact on either the velocity or turbulence intensity profiles. A slight decrease in both velocity and turbulence intensity was observed near the duct wall. The smaller the honeycomb, the lower the turbulence intensity, although the difference was minimal. At Re = 15000, the velocity profile was more symmetrical in ducts equipped with honeycomb straighteners. The turbulence intensity near the duct axis decreased by approximately 0.25 percentage points, regardless of honeycomb size.

At Re = 30000, the use of the honeycomb straightener influenced the velocity profile and turbulence intensity similarly across all honeycomb diameters. The velocity profile became more rounded, and turbulence intensity decreased by approximately 4 percentage points. Comparable results were obtained at Re = 45000. The measurement for the D5L20 was not included in Fig. 4d, as the experimental setup shown in Fig. 1 could not maintain a uniform flow for the duration required to capture the velocity profile data for this configuration.

3.4. Pressure drop on the straightener

Figure 5 shows the pressure drop measured over 1 meter of the duct where the straighteners were installed. The lowest pressure drop was recorded for the empty channel while the installation of any straightener resulted in increased hydraulic resistance. However, at Re = 10000, the drops were similar across all tested configurations. The lowest pressure drop was obtained for the D20L20 straightener. In general, a smaller honeycomb diameter corresponds to a higher pressure drop. The pressure drops for D10L20, D10L10, and D10L40 were identical, indicating that increasing the straightener length did not significantly affect the pressure loss. However, for the shortest straightener D10L5, higher losses were noted. These results indicate that, in addition to the typical pressure drop associated with the presence of shear stresses at the wall-air interface, straighteners can generate other disturbances that cause pressure drops. The research methodology used in this paper cannot indicate the nature of these phenomena. The highest pressure loss was shown for the D5L20 straightener, i.e., the straightener with the smallest diameter of the honeycomb.

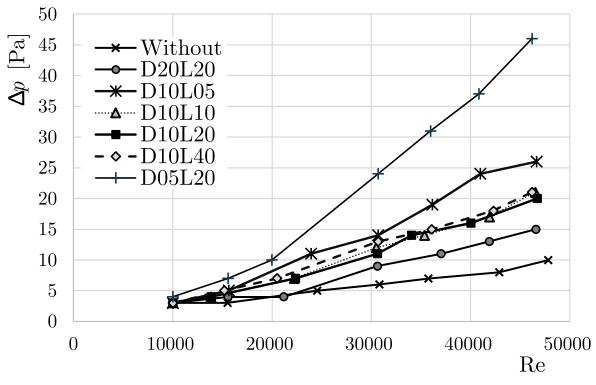

Fig. 5. Pressure drop in a honeycomb straightener for different Reynolds numbers.

4. Discussion

Figure 6 presents the relationship between axis turbulence intensity and the type of straighteners used. In all cases, the use of the honeycomb straightener resulted in a reduction in axial turbulence intensity. However, at low Reynolds numbers (Re = 10000 and Re = 15000), the reduction was minimal – 0.6 percentage points – regardless of the straightener configuration. Therefore, the use of flow straighteners at low Reynolds numbers appears unjustified. Despite their ineffectiveness, they introduce significant pressure losses in the system – exceeding 200 % at Re = 15000 for the D05D20 configuration (Fig. 5).

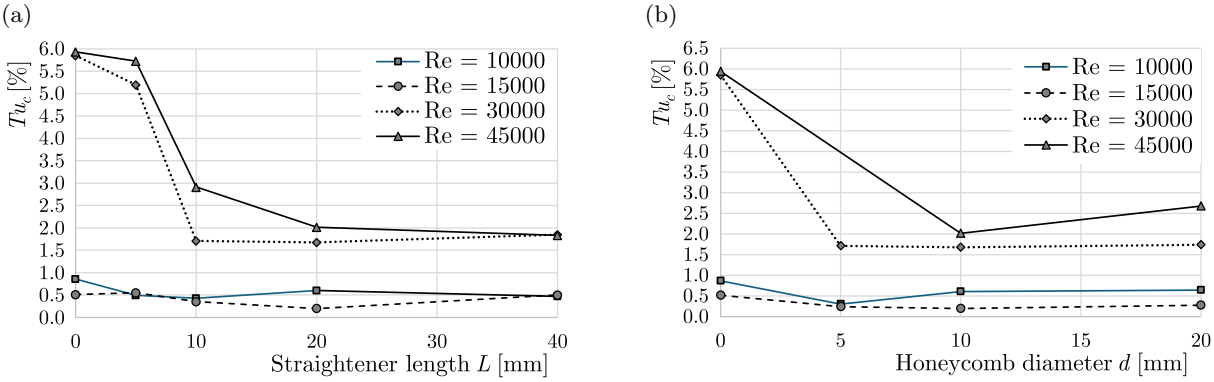

Fig. 6. Axis turbulence intensity for straighteners with different lengths (a) and honeycomb diameter (b).

At Re = 30000 and Re = 45000, the straighteners reduced turbulence intensity by more than 60 %. The reduction in turbulence increased with the straightener length up to 20 mm, after which it stabilized. The lack of impact of the stream straightener was observed from length 10 mm at Re = 30000. It should therefore be assumed that the higher the Reynolds numbers, the longer straighteners are required to achieve effective flow conditioning. The increase in honeycomb diameter did not affect the turbulence intensity at Re = 30000. However, when increasing the diameter to 20 mm, the axial turbulence also increased at Re = 45000. These findings indicate that smaller honeycomb diameters are more effective in reducing turbulence. It is important to note, however, that smaller honeycomb diameters also result in higher pressure losses (Fig. 5). Therefore, a honeycomb diameter of 10 mm appears to offer the best balance between performance and pressure drop.

4.1. Velocity profile analysis

A suitable method for quantifying the impact of straighteners on the velocity profile has not been established. Kühnen *et al.* (2018) investigated flow relaminarization at Re $\leq$ 6000, while Marensi *et al.* (2019) examined the impact of velocity profile flattening on drag reduction. The use of flow straighteners resulted in increased pressure losses (Fig. 5). However, the accurate measurement of pressure drop in the duct is challenging – especially at short distances – due to the need for highly precise instrumentation. Therefore, we propose an alternative approach based on evaluating the exponent n, calculated under the condition defined in Eq. (2.4). The values of exponent n derived from the measured velocity profiles are presented in Table 2. It is commonly assumed that $n = 7$ corresponds to fully developed turbulent flow (Salama, 2021). However, as shown in Table 2, the actual n values that best fit the experimental data are considerably higher. The values of n and the changes between them during the use of the straighteners only slightly reflect the observable changes in the shape of the velocity profiles. Therefore, Fig. 7 presents the percentage changes of the exponent n for the flow straighteners. Since the velocity profile is a function of the parameter $1/n$, the percentage change of parameter n presented in Fig. 7 was calculated as follows:

$$PCn = \frac{\frac{1}{n_i} - \frac{1}{n_{\text{without}}}}{\frac{1}{n_{\text{without}}}}, \tag{4.1}$$

where n_{without} is an exponent calculated for the velocity profile measured in the duct without straighteners, and n_i is an exponent calculated for the velocity profile measured in the duct with straighteners.

Table 2. Exponent n calculated for velocity profiles.

Reynolds number	Case						
	Without	D10L5	D10L10	D10L20	D10L40	D5L20	D20L20
10000	55.09	65.97	47.76	46.41	39.48	47.38	44.29
15000	50.97	54.94	37.22	37.17	34.52	34.92	21.25
30000	13.93	10.86	8.44	8.37	6.78	7.80	6.52
45000	9.50	8.47	5.28	5.28	5.26	–	5.28

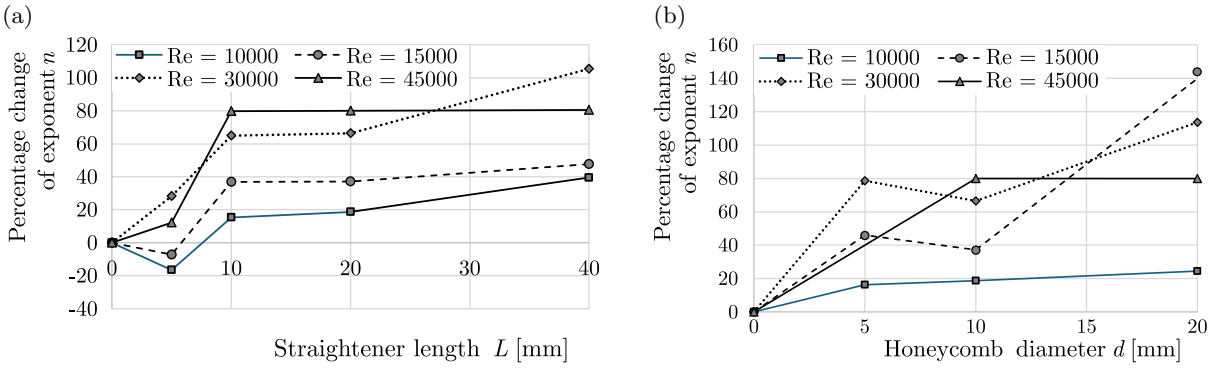

Fig. 7. Exponent n for different velocity profiles obtained for straighteners with different lengths (a) and honeycomb diameter (b).

The strongest impact of straighteners on the flow was observed at Re = 30000 and 45000. The percentage change in the exponent n was similar for all straighteners longer than 5 mm (Fig. 7a). In contrast, at Re = 10000 and 15000, the percentage change of the exponent n was negative – indicating that the value of the exponent n increased and the velocity profile became less laminar and more turbulent. These findings are consistent with the results reported by Xiong *et al.* (2003), who demonstrated that short straighteners (e.g., screens) initially cause an increase in turbulence level. For straighteners with different honeycomb diameters (Fig. 7b), an improvement in the velocity profile shape was observed in all cases. The highest percentage change in the exponent n occurred at Re = 15000, with a honeycomb diameter of $d = 20$ mm. Although this change is barely visible in the velocity profile (Fig. 4b), it is noteworthy that the use of the straightener reduced the n value from approximately 50.97 to 21.25. This is a large percentage change, but for high n values, it has little impact on the shape of the velocity profile.

The proposed method, based on the analysis of the exponent n, is relatively difficult to interpret. Therefore, an alternative approach is proposed, based on the kinetic energy correction factor, which can be calculated using the following formula:

$$\alpha = \frac{E_{\text{real}}}{E_{\text{ideal}}} = \frac{1}{A}\int_A \left(\frac{U(r)}{U_{\text{MEAN}}}\right)^3 \mathrm{d}A, \tag{4.2}$$

where E is the kinetic energy of the flow [J], A is a cross-section area of the duct $A = \frac{\pi D^2}{4}$ m^2, U_{MEAN} is a mean velocity in the duct [m/s].

Teleszewski (2018) showed that the kinetic energy correction factor is equal to 2 for the laminar flow, and approaches a value of 1 for fully turbulent flow. For this reason, it serves as a useful parameter for assessing changes in the shape of the velocity profile, which is directly related to the distribution of kinetic energy in the flow. Additionally, Teleszewski (2018) showed that the kinetic energy correction factor changes significantly within the Reynolds number range from 0 to 5000 ($\alpha(\mathrm{Re}=0)=2$ and $\alpha(\mathrm{Re}=5000)\approx 1.2$), but beyond this range the change is minimal $\alpha(\mathrm{Re}=20000)\approx 1.1$. Therefore, this parameter may be well-suited to evaluating the effect of a straightener on the flow profile. The calculated kinetic energy correction factors are presented in Fig. 8. For Reynolds numbers of 10000 and 15000, the use of a flow straightener had no noticeable effect on this parameter. This observation aligns with the velocity profile analysis, which also indicated minimal influence of the straighteners on the velocity profiles at these Reynolds numbers (Figs. 3a, 3b and Figs. 4a, 4b). The longer the straightener and the smaller the honeycomb diameter, the higher kinetic energy correction factor is observed at Re = 30000 and 45000. The largest increase in the parameter change is observed for the length of 10 mm and the diameter of 5 mm. At Re = 30000 and $d = 10$ mm, a decrease in the kinetic energy correction factor was observed.

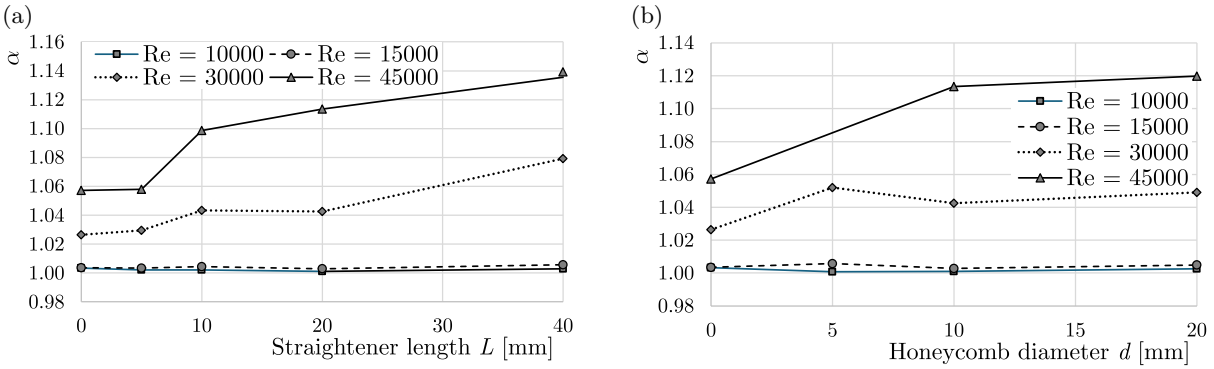

Fig. 8. Kinetic energy correction factor for different velocity profiles obtained for straighteners with different lengths (a) and honeycomb diameters (b).

An increase in the Reynolds number led to a decrease in the exponent n and an increase in the kinetic energy correction factor for cases without straighteners. While in a fully developed flow, the opposite trends would typically be observed and expected. However, in the present study, the flow is not fully developed, and shear stresses play a dominant role in altering the flow characteristics. As the Reynolds number increases, so does shear stress, resulting in a more rapid development of the velocity profile. This, in turn, influences the measured values of both the exponent n and the kinetic energy correction factor.

4.2. Minor loss coefficient of straighteners

Based on the measured pressure drop (Fig. 5), the major loss coefficient of the duct (Fig. 9a) and the minor loss coefficient of the straighteners (Fig. 9b) were calculated. The methodology for calculating loss coefficients is described in papers and several sources (Asker *et al.*, 2014). The minor loss coefficient of the straighteners was found to depend on the Reynolds number; however, no clear trend was observed. The highest minor loss coefficient was observed for the D05L20 straightener, while the lowest was for D20L20. The coefficients for D10L10, D10L20, and D10L40 straighteners were similar and, unexpectedly, lower than the shortest straightener D10L05. This suggests that the length of the straightener has a relatively minor influence on flow resistance. The D10L05 configuration may induce flow disturbances downstream, particularly at high Reynolds numbers, leading to additional losses. However, this hypothesis should be further validated using flow visualization techniques or particle image velocimetry (PIV). Among all

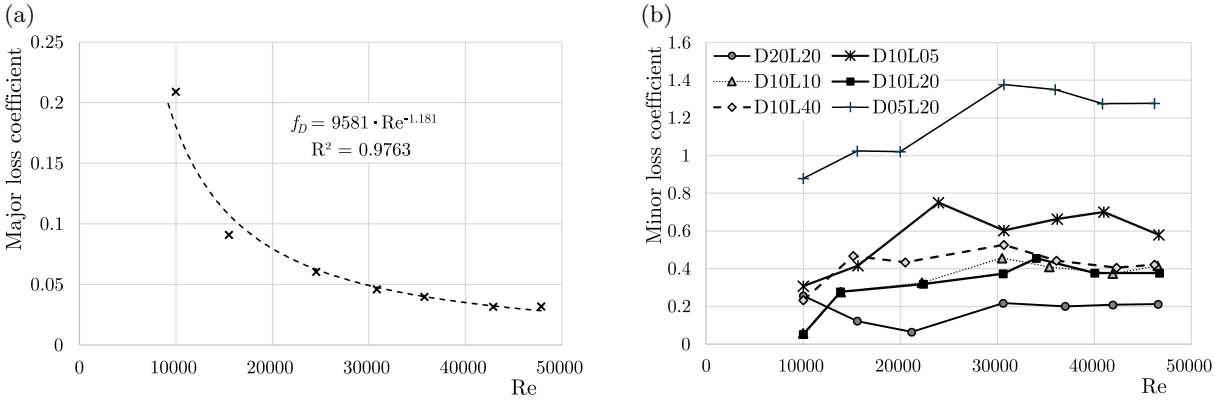

Fig. 9. Major loss factor in a straight duct (a) and the minor loss coefficient of straighteners (b).

parameters, the honeycomb diameter appears to exert the greatest influence on the minor loss coefficient.

It should be noted that the straightener reduces the cross-sectional flow area by introducing additional resistance surfaces – specifically, the front walls of the honeycomb cells. In the cases of D10L05, D10L10, D10L20, and D10L40, the reduction in flow area is identical, which explains the similar minor loss coefficients observed. The total wall surface area is the lowest for the D20L20 configuration and the highest for D05L20. This is due to maintaining the same wall thickness of 1 mm for all straighteners, rather than maintaining a constant flow area. Therefore, future research should include tests on straighteners with identical geometric dimensions (diameter and length) but varying wall thicknesses to better understand the influence of wall geometry on flow resistance.

5. Conclusions

The hot-wire anemometer was used to measure velocity profiles in a straight duct, both with and without a honeycomb straightener. Measurements were taken 1 meter downstream of the straightener and 2 meters downstream of the axial fan, which served as a turbulence generator. Additional measurements were also conducted at the fan outlet and 1 meter downstream of the fan. This study investigated the influence of the straightener length and honeycomb cell diameter on turbulence intensity, velocity profile, and pressure drop. Two methods were proposed to analyze the effect of the straightener on the velocity profile: the first is based on changes in the exponent n, while the second one utilizes the kinetic energy correction factor.

Based on the presented results and discussion, the following general conclusions can be drawn:

- the use of straighteners is primarily justified at high Reynolds numbers (Re = 30000, 45000), where the turbulence intensity is high and its decrease is most significant;
- straighteners influence the flow velocity profile at high Reynolds numbers, making it more parabolic and less rectangular – an effect referred to as flow relaminarization. The change in the profile shape can be quantified by analyzing the exponent n or the kinetic energy correction factor;
- even very short stream straighteners can reduce the level of flow turbulence, although as the straightener length increases, the reduction in turbulence intensity may increase (up to a certain point);
- as the honeycomb diameter increases, the straightener efficiency decreases;
- the honeycomb diameter has a greater influence on flow resistance than the straightener length. Notably, short straighteners ($l = 5$ mm) can induce higher pressure losses compared to longer ones.

The results presented in this study can serve as practical guidelines for the design of stream straighteners. The comprehensive dataset - including velocity and turbulence intensity profiles – can support further simulations aimed at extending this research. The analytical methods used to assess the velocity profile shape, such as the exponent n and kinetic energy correction factor, can also be compared with other approaches, including those based on major loss coefficient analysis. Finally, this study highlights promising directions for further research.

References

1. Asker, M., Turgut, O.E., & Coban, M.T. (2014). A review of non iterative friction factor correlations for the calculation of pressure drop in pipes. *Bitlis Eren University Journal of Science and Technology*, *4*(1), 1–8. https://dergipark.org.tr/tr/download/article-file/40279
2. Bradshaw, P. (1965). The effect of wind-tunnel screens on nominally two-dimensional boundary layers. *Journal of Fluid Mechanics*, *22*(4), 679–687. https://doi.org/10.1017/S0022112065001064
3. Dróżdż, A., Sokolenko, V., & Elsner, W. (2025). Performance analysis of novel wavy-wall-based flow control method for wind turbine blade. *Experimental Thermal and Fluid Science*, *169*, Article 111527. https://doi.org/10.1016/j.expthermflusci.2025.111527
4. Dutta, P., Rajendran, N.K., Cep, R., Kumar, R., Kumar, H., & Nirsanametla, Y. (2025). Numerical investigation of Dean vortex evolution in turbulent flow through 90° pipe bends. *Frontiers in Mechanical Engineering*, *11*. https://doi.org/10.3389/fmech.2025.1405148
5. El Drainy, Y.A., Saqr, K.M., Aly, H.S., & Jaafar, M.N.M. (2009). CFD analysis of incompressible turbulent swirling flow through Zanker plate. *Engineering Applications of Computational Fluid Mechanics*, *3*(4), 562–572. https://doi.org/10.1080/19942060.2009.11015291
6. Groth, J., & Johansson, A.V. (1988). Turbulence reduction by screens. *Journal of Fluid Mechanics*, *197*, 139–155. https://doi.org/10.1017/S0022112088003209
7. Hamzah, H., Jasim, L.M., Alkhabbaz, A., & Sahin, B. (2021). Role of honeycomb in improving subsonic wind tunnel flow quality: Numerical study based on orthogonal grid. *Journal of Mechanical Engineering Research and Developments*, *44*(7), 352–369.
8. Hrúz, M., Pecho, P., & Bugaj, M. (2020). Design procedure and honeycomb screen implementation to the Air Transport Department's subsonic wind tunnel. *AEROjournal*, *16*(2), 3–8. https://doi.org/10.26552/aer.C.2020.2.1
9. Hwang, Y. (2024). Near-wall streamwise turbulence intensity as $\mathrm{Re}_\tau \to \infty$. *Physical Review Fluids*, *9*(4), Article 044601. https://doi.org/10.1103/PhysRevFluids.9.044601
10. International Organization for Standardization. (2022). *Measurement of fluid flow by means of pressure differential devices inserted in circular cross-section conduits running full – Part 1: General principles and requirements* (ISO Standard No. 5167-1:2022). https://www.iso.org/standard/79179.html
11. Jurga, A.P., Janocha, M.J., Ong, M.C., & Yin, G. (2024). Numerical investigations of turbulent flow through a 90-degree pipe bend and honeycomb straightener. *Journal of Fluids Engineering*, *146*(2), Article 021307. https://doi.org/10.1115/1.4064101
12. Kaminski, K., Znaczko, P., Kardas-Cinal, E., Chamier-Gliszczynski, N., Koscielny, K., & Cur, K. (2025). Comparison of the heat transfer efficiency of selected counterflow air-to-air heat exchangers under unbalanced flow conditions. *Energies*, *18*(1), Article 117. https://doi.org/10.3390/en18010117
13. Klotz, L., Bukowski, K., & Gumowski, K. (2024). Influence of porous material on the flow behind a backward-facing step: experimental study. *Journal of Fluid Mechanics*, *998*, Article A31. https://doi.org/10.1017/jfm.2024.639
14. Kühnen, J., Scarselli, D., Schaner, M., & Hof, B. (2018). Relaminarization by steady modification of the streamwise velocity profile in a pipe. *Flow, Turbulence and Combustion*, *100*(4), 919–943. https://doi.org/10.1007/s10494-018-9896-4
15. Laws, E.M. (1990). Flow conditioning—A new development. *Flow Measurement and Instrumentation*, *1*(3), 165–170. https://doi.org/10.1016/0955-5986(90)90006-S

16. Lumley, J.L., & McMahon, J.F. (1967). Reducing water tunnel turbulence by means of a honeycomb. *Journal of Basic Engineering*, *89*(4), 764–770. https://doi.org/10.1115/1.3609700

17. Marensi, E., Willis, A.P., & Kerswell, R.R. (2019). Stabilisation and drag reduction of pipe flows by flattening the base profile. *Journal of Fluid Mechanics*, *863*, 850–875. https://doi.org/10.1017/jfm.2018.1012

18. Salama, A. (2021). Velocity profile representation for fully developed turbulent flows in pipes: A modified power law. *Fluids*, *6*(10), Article 369. https://doi.org/10.3390/fluids6100369

19. Saunders, G.P., Muller, S., Geierman, R., Duell, E., & Wagner, D.A. (2004). Wake effects of honeycomb stiffeners. *24th AIAA Aerodynamic Measurement Technology and Ground Testing Conference*. AIAA. https://doi.org/10.2514/6.2004-2199

20. Smyk, E., Stopel, M., & Szyca, M. (2024). Simulation of flow and pressure loss in the example of the elbow. *Water*, *16*(13), Article 1875. https://doi.org/10.3390/w16131875

21. Sun, K., Sun, J., Fan, Y., Yu, L., Chen, W., Kong, X., & Yu, C. (2023). Characterization of a synthetic jet vortex ring flowing through honeycomb. *Physics of Fluids*, *35*(7), Article 075123. https://doi.org/10.1063/5.0155935

22. Sun, K., Zhang, S., Shi, N., Peng, S., Cao, J., Sun, J., & Chen, W. (2025). Experimental investigation of synthetic jet impingement upon a honeycomb. *European Journal of Mechanics – B/Fluids*, *111*, 319–333. https://doi.org/10.1016/j.euromechflu.2025.02.003

23. Tan-Atichat, J., Nagib, H.M., & Loehrke, R.I. (1982). Interaction of free-stream turbulence with screens and grids: a balance between turbulence scales. *Journal of Fluid Mechanics*, *114*, 501–528. https://doi.org/10.1017/S0022112082000275

24. Teleszewski, J.T. (2018). Experimental investigation of the kinetic energy correction factor in pipe flow. *E3S Web of Conferences*, *44*, Article 00177. https://doi.org/10.1051/e3sconf/20184400177

25. Xiong, W., Kalkühler, K., & Merzkirch, W. (2003). Velocity and turbulence measurements downstream of flow conditioners. *Flow Measurement and Instrumentation*, *14*(6), 249–260. https://doi.org/10.1016/S0955-5986(03)00031-1

Manuscript received May 28, 2025; accepted for publication July 21, 2025; published online October 7, 2025.

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 795–807, Warsaw 2025
https://doi.org/10.15632/jtam-pl/209747

EFFECT OF FATIGUE ON THE MICROHARDNESS OF SCRAP CROSS-SECTIONS AFTER CYCLIC BENDING WITH TORSION OF RG7 BRONZE ALLOY

Mariusz PRAŻMOWSKI*, Joanna MAŁECKA, Tadeusz ŁAGODA

Mechanical Engineering, Opole University of Technology, Opole, Poland
*corresponding author, *m.prazmowski@po.edu.pl*

This work analyzes the effect of fatigue on the microhardness of the fracture plane of bronze samples. The analysis will be based on tests under conditions of cyclic bending, cyclic torsion, and two combinations of bending and torsion of samples made of RG7 bronze. All tests were performed at zero mean stress. The fracture plane was divided into a grid at 0.4 mm intervals, and local microhardness values were determined. On this basis, contour lines of microhardness were determined. The analysis of these contours on the surface showed that the most significant increase in the maximum microhardness in relation to the starting material was obtained for the static tension and cyclic bending tests. However, for the combination of cyclic bending and torsion, a minimal influence of shear stress in the maximum microhardness was obtained.

Keywords: fatigue life-time; bending with torsion; micro hardness; fracture plane.

1. Introduction

The relationship between strength and hardness of a material is well known. This also applies to the influence of material hardness on fatigue life. This influence can be found in numerous publications. Some papers also propose full formulas in which the fatigue strength (fatigue strength limit) is a function of hardness, or even, on this basis, full fatigue characteristics can be determined. The test results of the Ni alloy samples showed that as the size of the crystallites decreased, the experimental fatigue life of the samples increased (Sriraman *et al.*, 2007). This is due to the increase in the Vickers hardness (HV) value. In (Bandara *et al.*, 2016), the S-N fatigue characteristics in the range from small to gigacycles for steel with a tensile strength not exceeding 1400 MPa were derived and verified. This characteristic is an empirical characteristic based on the Brinell hardness (HB) of the analyzed steel and has the typical form for this range, i.e., the letter S as presented, among others, in (Kurek *et al.*, 2019). It was verified on the basis of fatigue tests with different cycle asymmetry coefficients for samples with and without a notch. In (Assi & Alkalali, 2021), the analysis of 5 steels and 5 aluminum alloys revealed that the fatigue limit in these two material groups depends linearly on HB.

Görzen *et al.* (2022) found that there is a linear relationship between the fatigue limit and HV, based on fatigue tests of five steels: X0.5CuNi2-2, X21CuNi2-2, 42CrMo4, 100CrMnSi6-4, and C50E.

Ministry of Science and Higher Education
Republic of Poland

This publication has been funded by the Polish Ministry of Science and Higher Education under the Excellent Science II programme "Support for scientific conferences".
The content of this article was presented during the 61st Symposium "Modelowanie w mechanice" (Modelling in Mechanics), Szczyrk, Poland, March 2–5, 2025.

Roessle & Fatemi (2000) proposed the deformation characteristics as a function of HB. This relationship is a quadratic function due to the hardness in the following form:

$$\frac{\Delta\varepsilon}{2} = \frac{4.25\,\mathrm{HB} + 225}{E}\,(2N_f)^{-0.09} + \frac{0.32\,\mathrm{HB}^2 - 487\,\mathrm{HB} + 191000}{E}\,(2N_f)^{-0.56}\,. \tag{1.1}$$

In (Shamsaei & Fatemi, 2009), for 1050 steel in three different states, which resulted in 3 different hardnesses (198, 360, and 565 HB) and different fatigue characteristics, an analogous expression to Eq. (1.1) was proposed, except that it is dependent on shear deformation on HB by modifications of the Fatemi–Socie model for multiaxial load condition:

$$\frac{\Delta\gamma_{\max}}{2}\left(1 + k\frac{\sigma_{n,\max}}{\sigma_y}\right) = \frac{6.37\,\mathrm{HB} + 338}{E}\,(2N_f)^{-0.09} + \frac{0.55\,\mathrm{HB}^2 - 842\,\mathrm{HB} + 331000}{E}\,(2N_f)^{-0.56}\,, \tag{1.2}$$

where

$$k = (0.0003\,\mathrm{HB} + 0.0585)\,(2N_f)^{0.09}\,. \tag{1.3}$$

Other methods have also been proposed, including those based on the ultimate tensile strength, i.e., σ_B:

– Mitchell model (Mitchell, 1996):

$$\sigma_B = 3.45\,\mathrm{HB}, \tag{1.4}$$

– Roessle–Fatemi model (Roessle & Fatemi, 2000):

$$\sigma_B = 0.0012\,\mathrm{HB}^2 + 3.3\,\mathrm{HB}, \tag{1.5}$$

– Baumel–Seeger recommendation and Kloos–Velten model (Kloos & Velten, 1984):

$$\sigma_B = 3.29\,\mathrm{HV} - 47 \quad \text{for} \quad \mathrm{HV} \leq 445, \tag{1.6}$$

$$\sigma_B = 4.02\,\mathrm{HV} - 374 \quad \text{for} \quad \mathrm{HV} < 445, \tag{1.7}$$

– method proposed in (Shiozawa & Sakai, 1996):

$$\sigma_B = \frac{\mathrm{HV} - 1.837}{0.304}. \tag{1.8}$$

Li *et al.* (2015) summarized the relationships between fatigue strength and ultimate stress as well as hardnesses expressed by HV, HB, and HRC (Rockwell hardness). The list of 14 linear or square relationships between fatigue limit and the same hardness was made on the basis of proposals from the literature on testing such materials as: steels and aluminum, copper, titanium, and magnesium alloys for the first relationship, and steel (Pang *et al.*, 2014) and Cu-Be alloy (Pang *et al.*, 2013) for hardness. The proposed dependencies in the hardness function are simple, linear or square mathematical functions of the hardness of the analyzed materials.

James *et al.* (2009) found that hot spot strain in a welded joint is a linear function of HV versus residual strain. This, in turn, has a linear (double-logarithmic) effect on the experimental life time.

Xin *et al.* (2021) proposed the four-parameter Bandara stress fatigue characteristic for welded joints (Bandara *et al.*, 2015; 2016):

$$\sigma_a = a(N_f + B)^b + c, \tag{1.9}$$

where the coefficients a, B, and c were determined based on various physical quantities, including the HV of the tested material. The exponent b was assumed as constant, equal to -0.20.

Kondo *et al.* (2003), based on measurements and analyses, showed that the stress intensity factor is linearly dependent on the HV according to the following equation:

$$\Delta K_{th} = 3.3 \cdot 10^{-3}(\mathrm{HV} + 120)\left(\sqrt{\mathrm{area}}\right)^{1/3}. \tag{1.10}$$

The exceptions in the literature are two papers examining the effect of fatigue on the hardness of the material. The first work is Pavlou's (2002) research. Based on the tests of aluminum alloy 2024-T42, a linear dependence of the increase in HV was shown, depending on the number of cycles (n) for the given cycle amplitude. It has been shown that the degree of damage is a function of HV, depending on the number of cycles, i.e., stress:

$$D(n, \sigma) = \mathrm{HV}(n, \sigma). \tag{1.11}$$

The second paper examining the effect of fatigue on the hardness of the material was written by Rogachev *et al.* (2023) on the effect of the Cu–Zn alloy material. In this case, a sheet of this bronze with a thickness of 3 mm has undergone technological alternating bending. As a result of such a process, deformations in the elastic-plastic range changed in the processed element. The starting material had a microhardness of 99 HV. After the technology used, the average cross-sectional hardness increased to 124 HV. The greatest strengthening, to the level of 130 HV, was observed on the outer surfaces, where the greatest deformations occurred, and the smallest in the central part. There, the hardness increased to 118 HV.

No more papers analyzing the effect of fatigue on hardness have been found. It seems that the task opposite to what was presented in the review of the literature on the problem under consideration may also be interesting from the cognitive point of view. During the fatigue process, the material undergoes deformation, and significant plastic deformations occur locally. These deformations can determine the microhardness variables. This process, in the case of tension-compression, may be less interesting due to the homogeneous nature of both strain and stress. However, in the case of stress and strain gradients, this process can be particularly noticeable. It appears that the simplest fatigue tests, followed by microhardness analysis, can be conducted on the basis of tests under conditions of cyclic bending, cyclic torsion, and a combination of cyclic bending and torsion. At the same time, a fractographic and topographical analysis of the fractures obtained should be carried out. In this way, we will get a picture of the hardness and surface quality for different combinations. It seems that such an image may define a previous load. Therefore, the resulting image in the combination of topography and hardness determines the previous fatigue load.

The aim of this work is to analyze the effect of fatigue on samples made of RG7 bronze on the microhardness on the fracture plane with cracks. The analysis was performed on the basis of cyclic bending, cyclic torsion, and two combinations of proportional bending and torsion at zero mean stress.

2. Experimental research

Experimental studies concern the RG7 copper alloy (other designations are, for example, CuSn7Zn4Pb7, CC993K), where the elastic modulus $E = 92.14$ GPa. This material is characterized by very high ductility (Hong, 2018; Lim *et al.*, 2009; You & Miskiewicz, 2008; Małecka *et al.*, 2023), like most materials where the main component is copper. The static properties of the tested and analyzed bronze are characterized by the yield point $\sigma_y = 120$ MPa, ultimate stress $\sigma_u = 270$ MPa (Małecka & Łagoda, 2024).

Fatigue tests will be performed on samples without a geometric notch of the "diabolo" type (Fig. 1) for pure symmetrical plane bending, pure double-sided torsion, and two combinations

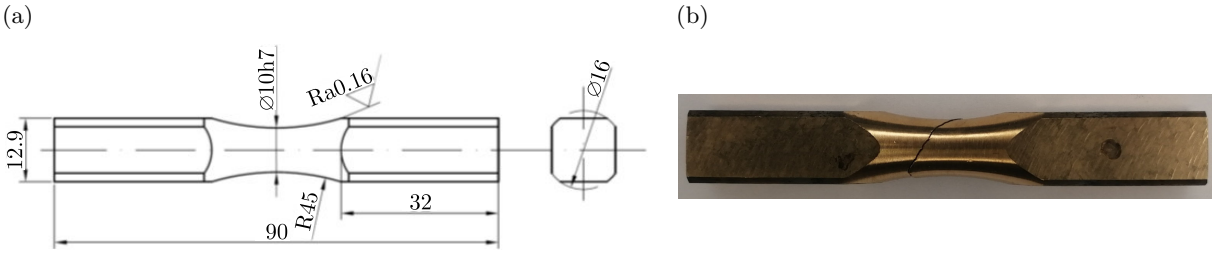

Fig. 1. "Diabolo" type specimen's dimensions: (a) working drawing; (b) photo of the sample after fatigue tests.

of proportional bending with torsion in relation to the amplitudes of stresses from torsion and bending, 0.5 and 1. This means:

$$\tau_a = 0.5\sigma_a, \qquad \tau_a = \sigma_a. \tag{2.1}$$

For the fatigue tests, a stand designed and made at the Opole University of Technology was used. Details of the operation of this stand can be found, among others, in (Małecka & Łagoda, 2024). The main principle of operation of this station is to spin the unbalanced mass, which in turn gives strength. This force, on the other hand, acts on the arm and results in a cyclically varying torque. This moment can, in effect, be split into any combination of alternating bending and torsional moments. Assuming elasticity, it can be stated that the tests are carried out under stress control. This allows the testing of materials for any combination of proportional cyclic bending and torsion.

Fatigue tests will be carried out so that the minimum fatigue life is at a high load level, giving a maximum of approx. 50000 cycles and a minimum load of at least 1000000 cycles (close to the fatigue limit). On this basis, Basquin fatigue characteristics (based on 18–20 specimens for every characteristic) written in double-logarithmic scale will be determined for each material in the form of

$$\log N_f = A_\sigma - m_\sigma \log \sigma_a \tag{2.2}$$

or

$$\log N_f = A_\tau - m_\tau \log \tau_a. \tag{2.3}$$

In the case of cyclic bending and two combinations of cyclic bending and torsion, according to Eqs. (2.1), the coefficients appearing in Eq. (2.2) are as follows: $A_\sigma = 26.26$, $m_\sigma = 9.09$ for cyclic bending; $A_\sigma = 24.47$, $m_\sigma = 8.85$, and $A_\sigma = 25.24$, $m_\sigma = 10.64$ for a combination of cyclic bending and torsion.

In the case of torsion, the coefficients in Eq. (2.3) are: $A_\tau = 38.34$ and $m_\tau = 15.38$.

The nanohardness tester (Fig. 2a) (Derda *et al.*, 2022), and the distribution of microhardness in the cross-section of the damaged material will be presented. Additionally, an optical microscope was used (Fig. 2b). Static and uniaxial tensile-compression tests were performed on a standard INSTRON fatigue stand.

An important characteristic of construction materials is their hardness, which depends on properties such as ductility, stiffness, plasticity, deformability, and strength of the tested material. As part of the research, hardness distribution contour lines will be determined on the cross-sections of samples, both in the initial state and after being subjected to fatigue tests. The assessment of changes in the properties of the tested materials will be carried out based on Martens hardness measurements, which are a hardness testing method based on continuous measurement of force as a function of displacement. Unlike standard methods, which include

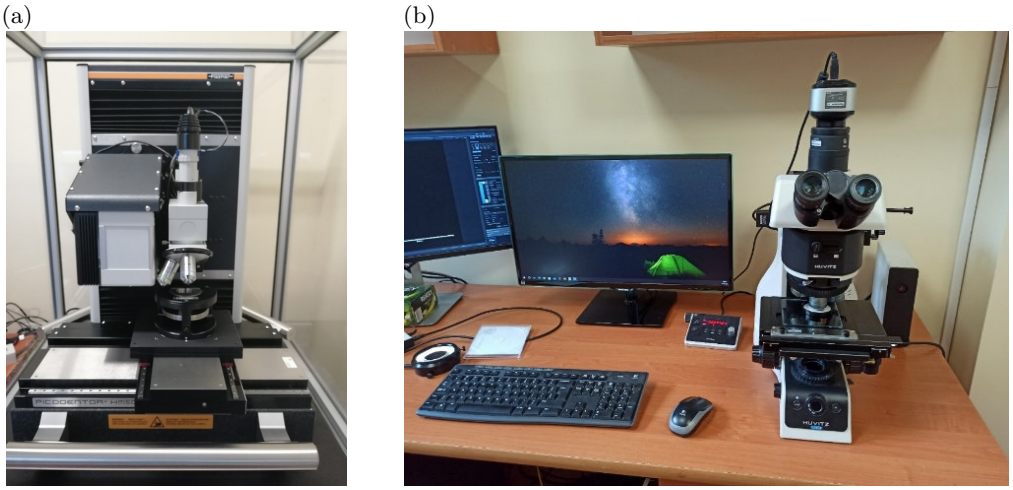

Fig. 2. Research devices: (a) nanohardness stand PICODENTOR HM 500 produced by Helmut Fischer; (b) optical microscope.

the Brinell, Rockwell, and Vickers methods, it is not based on a hardness reading from the measurement of the surface area of the indentation formed under the influence of a given force, but on a computer analysis of the obtained penetration curve. The measurement result is presented in the form of a loading and unloading curve as a function of the force applied to the indenter from the depth of penetration. Based on the obtained results, it will be possible to determine the stiffness of the sample read from the loading curve, the instrumental modulus of elasticity (approximately equal to the modulus of elasticity of the material), instrumental hardness, work of deformation (energy of elastic and plastic deformation of the material), etc. The tests will be performed using a system for measurement of nano- and microhardness according to the Martens method, in accordance with the PN EN ISO 14577 standard, PICODENTOR HM 500 (Fig. 2a) equipped with WIN-HCU software, which allows the use of forces applied to the indenter in the range of 0.005 mN–500 mN.

3. Microhardness measurements

Hardness measurements were made on selected (best suited to the fatigue characteristics) samples according to Table 1. For various bending-torsion combinations, samples (23, 44, 58, 21) were selected for a durability of approximately 500000 cycles. In addition, the sample was analyzed and not subjected to any load in the conditions of uniaxial static stretching and cyclic stretching-compression.

Table 1. Summary of sample numbers, number of cycles to failure and load method.

Sample	Number	N_f (cycles)
Withouth loading	00	–
Static tension	02	0.5
Tension-compression	06	5924
$\tau_a = 0$	23	571257
$\sigma_a = 0.5\tau_a$	44	776838
$\sigma_a = \tau_a$	58	522700
$\sigma_a = 0$	21	481710

Metallographic sections for macroscopic examinations and microhardness measurements were made from samples after fatigue tests, and the test surface was a cross-section perpendicular to

the axis of the sample at a distance of up to 3 mm from the obtained fracture. The samples were cut on a disc cutter with intensive cooling, and then embedded in a plastic mass. The samples prepared in this way were ground manually on abrasive papers of decreasing gradation (#350, #600, #800, #1200, #2000) and then mechanically polished on synthetic cloths using a water suspension of aluminum oxide (Al_2O_3). Finally, the samples were polished on a vibratory polisher and chemically etched with a reagent for etching copper and its alloys ($HCl+FeCl_3+H_2O$) to eliminate the effect of surface strengthening after metallographic preparation. The prepared microsection is shown in Fig. 3a and Fig. 4a.

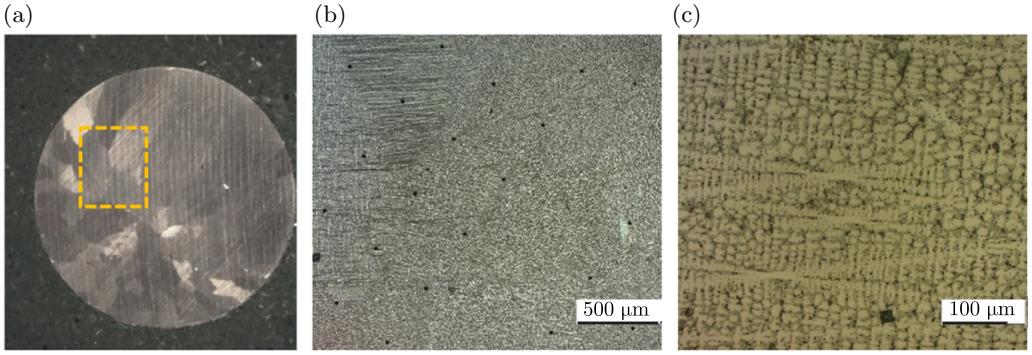

Fig. 3. Bronze sample in the initial state: (a) macro image with the micro observation point marked; (b), (c) bronze microstructure over 50× and 200×.

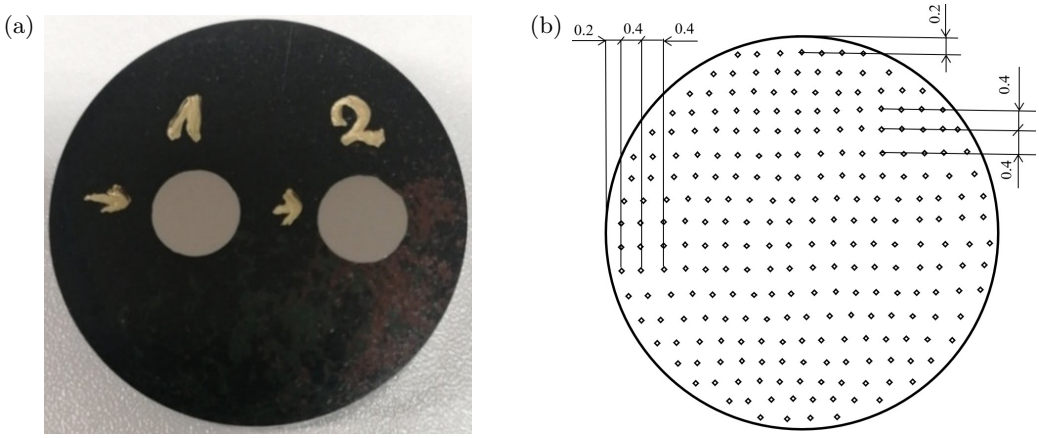

Fig. 4. Sample for microhardness testing: (a) macro-photo of the micro-section before microhardness measurement; (b) measurement scheme (500 to 600 points depending on the sample).

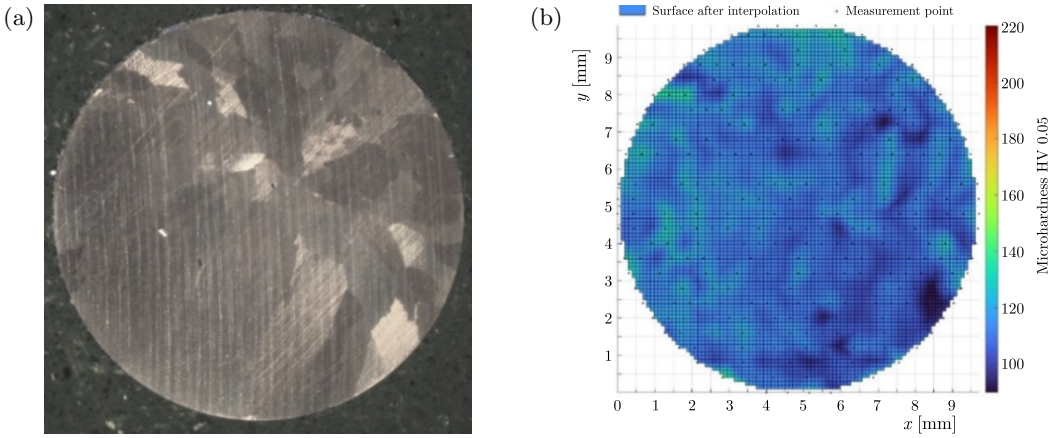

Fig. 5. Sample before loading (00): (a) macroscopic photo of the micro-section surface before microhardness measurement; (b) microhardness contour lines on the surface of the micro-section.

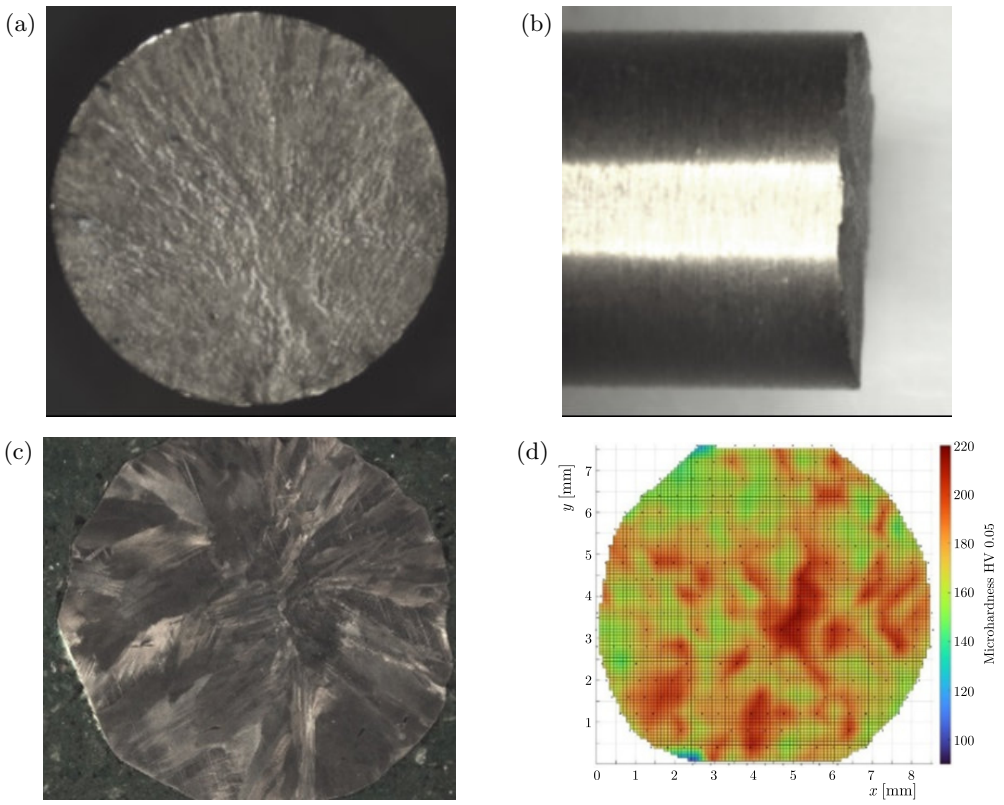

Fig. 6. Sample (02) after the static tensile test: (a) macroscopic photo of the scrap surface; (b) macroscopic photo of the crack direction; (c) macroscopic photo of the micro-section surface before microhardness measurement; (d) microhardness contour lines on the surface.

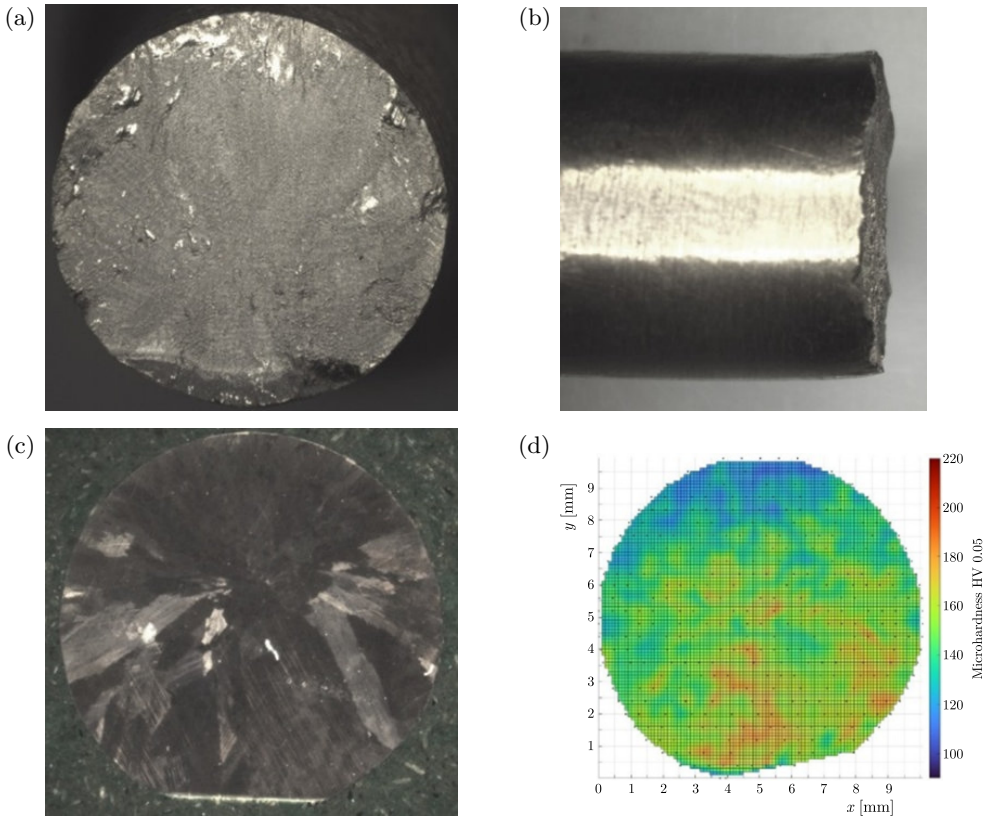

Fig. 7. Sample (06) after the tensile-compression test: (a) macroscopic photo of the scrap surface; (b) macroscopic photo of the crack direction; (c) macroscopic photo of the micro-section surface before microhardness measurement; (d) microhardness contour lines on the surface.

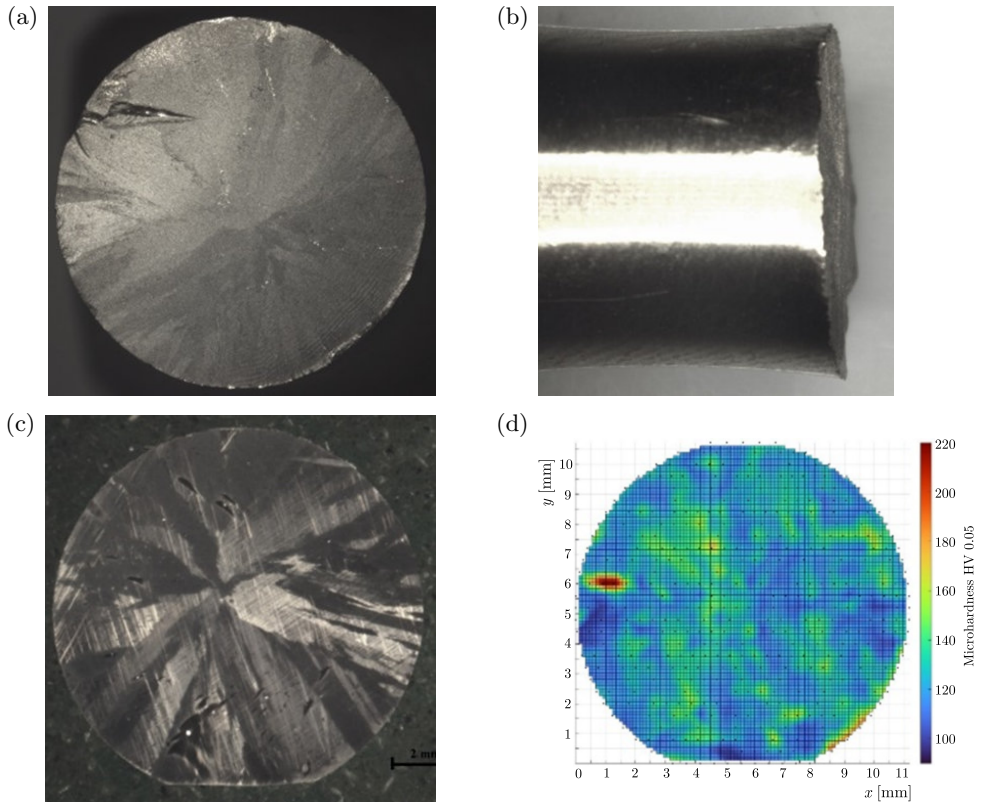

Fig. 8. Sample (23) after the cyclic bending test: (a) macroscopic photo of the scrap surface; (b) macroscopic photo of the crack direction; (c) macroscopic photo of the micro-section surface before microhardness measurement; (d) microhardness contour lines on the surface.

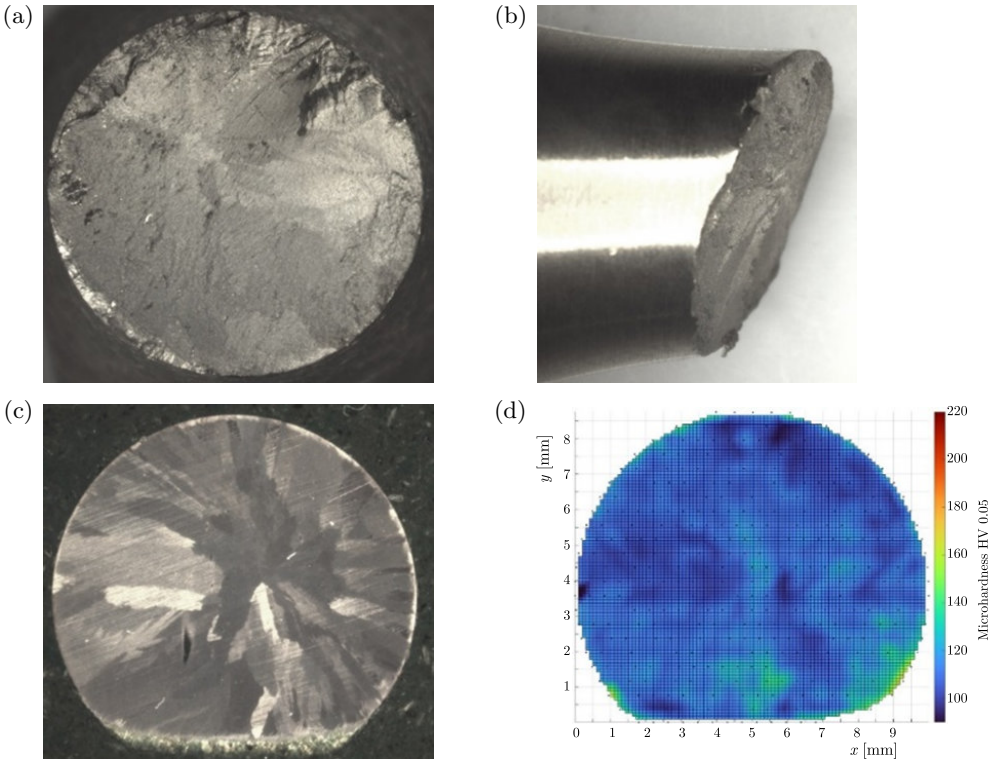

Fig. 9. Sample (44) after cyclic bending with torsion test $\sigma_a = 0.5\tau_a$: (a) macroscopic photo of the scrap surface; (b) macroscopic photo of the scrap surface before microhardness measurement; (c) microhardness contour lines on the surface.

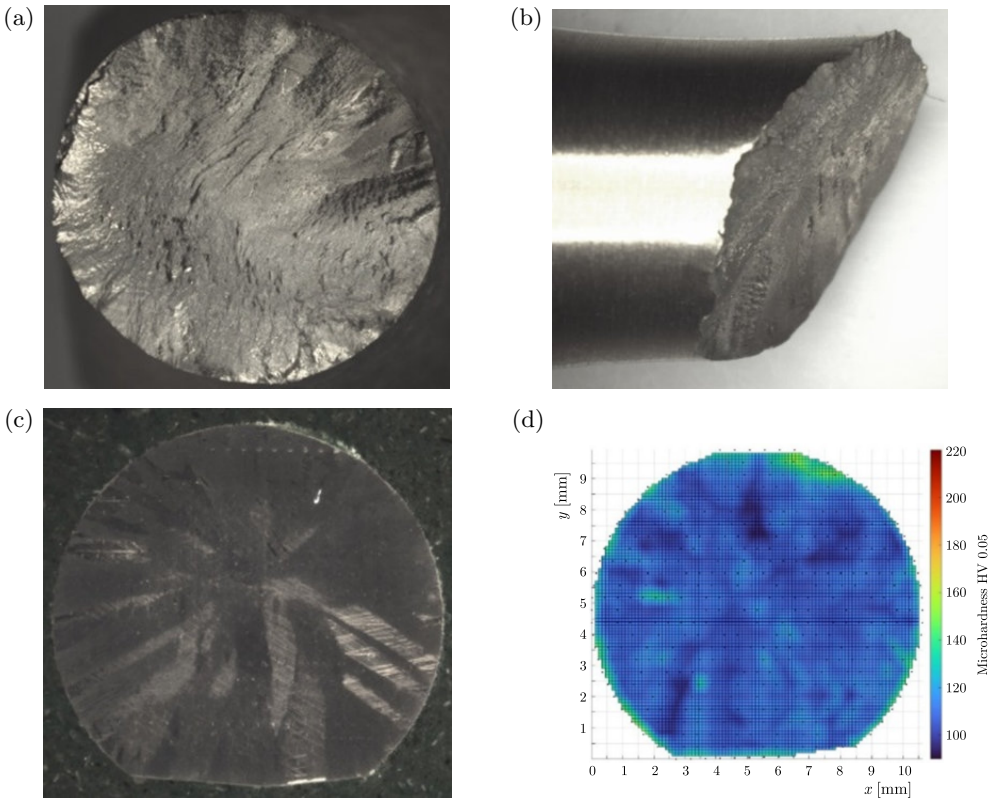

Fig. 10. Sample (58) after the cyclic bending test with torsion $\sigma_a = \tau_a$: (a) macroscopic photo of the scrap surface; (b) macroscopic photo of the crack direction; (c) macroscopic photo of the micro-section surface before microhardness measurement; (d) microhardness contour lines on the surface.

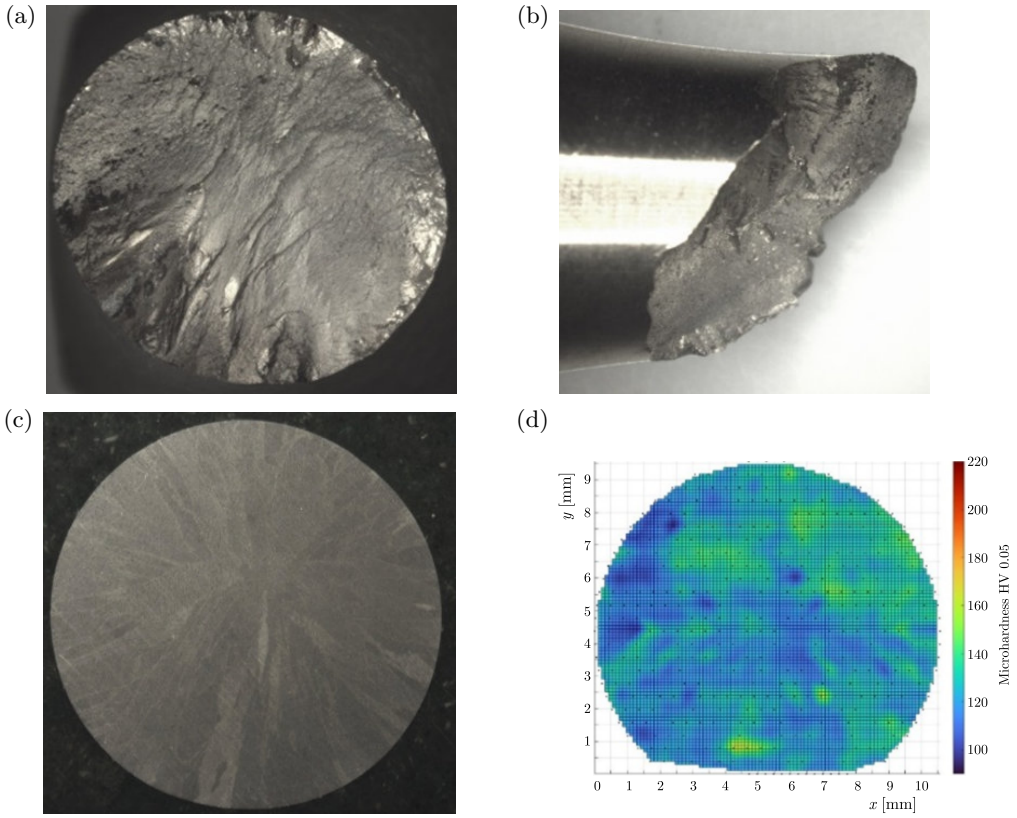

Fig. 11. Sample (21) after the cyclic torsion test: (a) macroscopic photo of the scrap surface; (b) macroscopic photo of the crack direction; (c) macroscopic photo of the micro-section surface before microhardness measurement; (d) microhardness contour lines on the surface.

The structure of the analyzed samples is typical for tin-lead bronzes, in which the α solid solution with the eutectoid phase $(\alpha + \delta)$ occurring in the interdendritic spaces and lead precipitates can be identified (Figs. 3b and 3c). Microhardness measurements were carried out using the Martens method, for which a Vickers indenter was used, which was loaded with a force of 500 mN (50 g) for 20 seconds. Measurement points were made as described in Fig. 4b.

Figure 5b and part (d) of Figs. 6–11 show the contour lines of the microhardness distribution for individual samples, i.e., the initial state (00), after the static tensile test (02), after the tensile-compression fatigue test (06), after cyclic bending (23), cyclic torsion (21), a combination of bending and torsion with a proportionality factor of 0.5 (44) and a proportionality factor of 1 (58) in combination with the image of the obtained fractures (part (a) of Figs. 5–11) along with the direction of the crack (part (b) of Figs. 6–11) and macrostructure (Fig. 5a and part (c) of Figs. 6–11). Pictures were taken with an optical microscope at a magnification of about 20×. In the case of cyclic bending (Fig. 8), a significant local increase in hardness can be observed, which is probably due to static fracture at this location.

Figure 12 shows the minimum, average, and maximum values of microhardness obtained for individual samples.

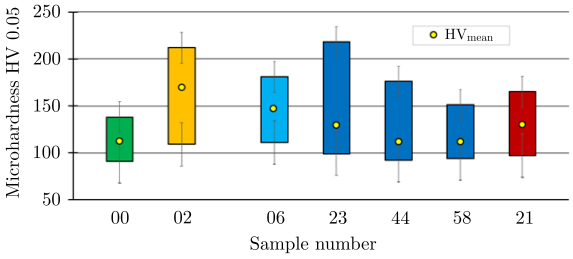

Fig. 12. List of microhardness for individual tested samples.

4. Analysis of the obtained measurement results

Table 2 presents the average microhardness values of the tested samples, taking into account the minimum and maximum values in the analyzed areas, shown in Fig. 5b and part (d) of Figs. 6–11. In addition, the maximum increase in microhardness in relation to the maximum microhardness of the unloaded sample was summarized. The maximum increase was obtained for the static tensile and cyclic bending tests. Then the deformations were the greatest, and locally the most significant hardness increase occurred. A relatively large increase in microhardness also occurred in the case of stretching and the combination of cyclic bending and torsion with small shear stresses. The smallest increase in maximum microhardness occurred in the case of a combination of bending and torsion with significant shear stresses.

Table 2. List of minimum, maximum, and medium microhardnesses HV.

Sample		HV$_{min}$	HV$_{max}$	HV$_{mean}$	Max increase HV$_{max}$ [%]
00	Without loading	91	138	114	–
02	Static tension	109	212	174	54
06	Tension-compression	111	181	148	31
23	Bending $\tau_a = 0$	99	218	126	57
44	$\sigma_a = 0.5 \cdot \tau_a$	92	176	113	28
58	$\sigma_a = \tau_a$	94	151	113	9
21	Torsion $\sigma_a = 0$	97	165	124	20

In the case of sample 00 (no load condition – starting material), the hardness of the sample oscillated in the range of 91 HV–138 HV and was clearly related to the heterogeneity of the

material structure, which can be seen in Fig. 5b. The largest increase in the average hardness in relation to the initial sample (00) (114 HV), which is 53 % (174 HV), was recorded for the sample after the static tensile test (02). Additionally, for this sample, the maximum hardness of 212 HV was obtained, i.e., an 85 % increase compared to the average hardness of the sample in the initial state, and it was located mainly in the central part of the cross-section. In the case of the sample that was subjected to the cyclic tensile test with compression (06) and subjected to bending (23), a systematic decrease in average hardness was observed, to the respective values of 148 HV and 126 HV, but these values are higher than the hardness of the initial sample by 30 % and 10 %, respectively. In all cases, the fracture formed was perpendicular to the axis of the sample (part (b) of Figs. 6–8), and the maximum hardness was identified in the areas where cracks were initiated (part (a) of Figs. 6–8).

The samples after the bending and torsion tests carried out at different loads (44 and 58) were characterized by an average hardness at the same level as in the case of the sample in the initial state (00) and both amounted to 113 HV. However, the maximum hardness in the analyzed areas was 176 HV and 151 HV, respectively, and it was identified near the edge of the sample (Fig. 9d and Fig. 10d). In the last case, i.e., the bending sample (21), there was a slight, 9 % increase in the average cross-sectional hardness (124 HV); however, the maximum hardness, at the level of 165 HV, was identified in different cross-sectional areas (Fig. 11d). A different nature of the resulting breakthroughs was also observed. In the case of samples subjected to bending with torsion, the fracture was formed at an angle of approx. 70°–75° to the axis of the sample, while in the case of torsion, this angle is approx. 50°.

5. Conclusions

The measurement and analysis of the microhardness of the RG7 bronze fracture plane showed that:

- every static or fatigue damage causes an increase in microhardness in relation to the unloaded material;
- the greatest increase in microhardness occurs at the point of the greatest unloading in the presence of a stress gradient;
- the highest increase in maximum microhardness was obtained for the static tensile test and for cyclic bending (in these cases, we are dealing with the largest surface on which destruction occurs at the initiation stage);
- the smallest increase in maximum microhardness was obtained for the combination of cyclic bending and torsion with a significant shear stress coming from torsion.

Acknowledgments

This work was financially supported by the Opole University of Technology as part of the GRAS project no. 260/23.

References

1. Assi, A.D., & Alkalali, R.H.M. (2021). Fatigue limit prediction based on hardness for both steel and aluminum alloys. *IOP Conference Series: Materials Science and Engineering*, *1105*, Article 012044. https://doi.org/10.1088/1757-899X/1105/1/012044

2. Bandara, C.S., Siriwardane, S.C., Dissanayake, U.I., & Dissanayake, R. (2015). Developing a full range S-N curve and estimating cumulative fatigue damage of steel elements. *Computational Materials Science*, *96* (Part A), 96–101. https://doi.org/10.1016/j.commatsci.2014.09.009

3. Bandara, C.S., Siriwardane, S.C., Dissanayake, U.I., & Dissanayake, R. (2016). Full range S–N curves for fatigue life evaluation of steels using hardness measurements. *International Journal of Fatigue*, *82*(Part 2), 325–331. https://doi.org/10.1016/j.ijfatigue.2015.03.021

4. Derda, S., Karolczuk, A., Prażmowski, M., Kurek, A., Wachowski, M., & Paul, H. (2022). Fatigue life and cyclic creep of tantalum/copper/steel layerwise plates under tension loading at room temperature. *International Journal of Fatigue*, *162*, Article 106977. https://doi.org/10.1016/j.ijfatigue.2022.106977

5. Görzen, D., Ostermayer, P., Lehner, P., Blinn, B., Eifler, D., & Beck, T. (2022). A new approach to estimate the fatigue limit of steels based on conventional and cyclic indentation testing. *Metals*, *12*(7), Article 1066. https://doi.org/10.3390/met12071066

6. Hong, S.I. (2018). Criteria for predicting twin-induced plasticity in solid solution copper alloys. *Materials Science and Engineering: A*, *711*, 492–497. https://doi.org/10.1016/j.msea.2017.11.076

7. James, M.N., Ting, S.-P., Bosi, M., Lombard, H., & Hattingh, D.G. (2009). Residual strain and hardness as predictors of the fatigue ranking of steel welds. *International Journal of Fatigue*, *31*(8–9), 1366–1377. https://doi.org/10.1016/j.ijfatigue.2009.03.006

8. Kloos, K.H., & Velten, E. (1984). Calculation of the fatigue strength of plasma nitrided component-like samples taking into account the hardness and residual stress profile (in German). *Konstruktion*, *36*(5), 181–8.

9. Kondo, Y., Sakae, C., Kubota, M., & Kudou, T. (2003). The effect of material hardness and mean stress on the fatigue limit of steels containing small defects. *Fatigue & Fracture of Engineering Materials & Structures*, *26*(8), 675–682. https://doi.org/10.1046/j.1460-2695.2003.00656.x

10. Kurek, A., Kurek, M., & Łagoda, T. (2019). Stress-life curve for high and low cycle fatigue. *Journal of Theoretical and Applied Mechanics*, *57*(3), 677–684. http://doi.org/10.15632/jtam-pl/110126

11. Li, Z., Wang, Q., Luo, A.A., Fu, P., & Peng, L. (2015). Fatigue strength dependence on the ultimate tensile strength and hardness in magnesium alloys. *International Journal of Fatigue*, *80*, 468–476. https://doi.org/10.1016/j.ijfatigue.2015.07.001

12. Lim, C.-B., Kim, K.S., & Seong, J.B. (2009). Ratcheting and fatigue behavior of a copper alloy under uniaxial cyclic loading with mean stress. *International Journal of Fatigue*, *31*(3), 501–507. https://doi.org/10.1016/j.ijfatigue.2008.04.008

13. Małecka, J., & Łagoda, T. (2024). Use of the biaxial coefficient in determining life for a combination of cyclic bending and torsion of bronze RG7. *Journal of Theoretical and Applied Mechanics*, *62*(3), 547–560. https://doi.org/10.15632/jtam-pl/188855

14. Małecka, J., Łagoda, T., Głowacka, K., & Vantadori, S. (2023). Influence of plastic deformations on both yield strength and torsional fatigue life of non-ferrous alloys. *Fatigue & Fracture of Engineering Materials & Structures*, *46*(6), 2080–2095. https://doi.org/10.1111/ffe.13982

15. Mitchell, M.R. (1996). Fundamentals of modern fatigue analysis for design. In ASM Handbook Committee (Eds.), *ASM Handbook: Vol. 19. Fatigue and Fracture* (pp. 227–249). ASM International. https://doi.org/10.31399/asm.hb.v19.a0002364

16. Pang, J.C., Li, S.X., Wang, Z.G., & Zhang, Z.F. (2014). Relations between fatigue strength and other mechanical properties of metallic materials. *Fatigue & Fracture of Engineering Materials & Structures*, *37*(9), 958–976. https://doi.org/10.1111/ffe.12158

17. Pang, J.C., Li, S.X., & Zhang, Z.F. (2013). High-cycle fatigue and fracture behaviours of Cu-Be alloy with a wide strength range. *Fatigue & Fracture of Engineering Materials & Structures*, *36*(2), 168–176. https://doi.org/10.1111/j.1460-2695.2012.01710.x

18. Pavlou, D.G. (2002). A phenomenological fatigue damage accumulation rule based on hardness increasing, for the 2024-T42 aluminum. *Engineering Structures*, *24*(11), 1363–1368. https://doi.org/10.1016/S0141-0296(02)00055-X

19. Roessle, M.L., & Fatemi, A. (2000). Strain-controlled fatigue properties of steels and some simple approximations. *International Journal of Fatigue*, *22*(6), 495–511. https://doi.org/10.1016/S0142-1123(00)00026-8

20. Rogachev, S.O., Shelest, A.E., Perkas, M.M., Andreev, V.A., Tabachkova, N.Yu., Yusupov, V.S., Ten, D.V., Isaenkova, M.G., & Krymskaya, O.A. (2023). Effect of alternating bending on structure, texture, and mechanical properties of Cu–Zn alloy. *Journal of Materials Engineering and Performance*, *33*(3), 1241–1249. https://doi.org/10.1007/s11665-023-08050-w

21. Shamsaei, N., & Fatemi, A. (2009). Effect of hardness on multiaxial fatigue behaviour and some simple approximations for steels. *Fatigue & Fracture of Engineering Materials & Structures*, *32*(8), 631–646. https://doi.org/10.1111/j.1460-2695.2009.01369.x

22. Shiozawa, K., Sakai, T. et al. (1996). *Databook on fatigue strength of metallic materials* (Vols. 1–3). Elsevier & JSMS.

23. Sriraman, K.R., Raman, S.G.S., & Seshadri, S.K. (2007). Influence of crystallite size on the hardness and fatigue life of steel samples coated with electrodeposited nanocrystalline Ni–W alloys. *Materials Letters*, *61*(3), 715–718. http://doi.org/10.1016/j.matlet.2006.05.049

24. Xin, H., Correia, J.A.F.O., Veljkovic, M., Berto, F., & Manuel, L. (2021). Residual stress effects on fatigue life prediction using hardness measurements for butt-welded joints made of high strength steels. *International Journal of Fatigue*, *147*, Article 106175. https://doi.org/10.1016/j.ijfatigue.2021.106175

25. You, J.-H., & Miskiewicz, M. (2008). Material parameters of copper and CuCrZr alloy for cyclic plasticity at elevated temperatures. *Journal of Nuclear Materials*, *373*(1–3), 269–274. https://doi.org/10.1016/j.jnucmat.2007.06.005

Manuscript received May 19, 2025; accepted for publication August 20, 2025; published online October 30, 2025.

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 809–814, Warsaw 2025
https://doi.org/10.15632/jtam-pl/210426

RESEARCH ON THE ENERGY RECOVERY SYSTEM IN THE ACTIVE HORIZONTAL SEAT SUSPENSION OF A WORKING MACHINE

Bartosz JERECZEK*, Igor MACIEJEWSKI, Andrzej BŁAŻEJEWSKI,
Sebastian PECOLT, Tomasz KRZYŻYŃSKI

Faculty of Mechanical and Energy Engineering, Koszalin University of Technology, Koszalin, Poland
*corresponding author, *bartosz.jereczek@tu.koszalin.pl*

This paper investigates the potential for longitudinal vibration energy recovery in a seat suspension system through the implementation of a brushless direct current (BLDC) motor. The work focuses on two states of system operation. The first one is when an electric motor works as an actuator in the powering mode to withstand horizontal forces. The second one occurs in the regenerative mode when the kinetic energy of the system is partially converted into electricity. Within the scope of the presented study, the efficiency of the energy regeneration process under random vibration conditions of different intensities is investigated experimentally. Measurement results are presented in the form of vibration transmittance functions for a suspension with energy regeneration compared to a conventional passive and fully active system.

Keywords: vibrations; seat suspension; active system; energy harvesting; recuperative braking.

1. Introduction

One of the main areas of current studies in mechanical and electrical engineering is energy harvesting from vibrating systems. The process by which mechanical energy produced by system or ambient vibrations is transformed into useable electrical energy is the basis for the phenomenon. It can be utilized to lower the system's overall energy consumption or to power (Paul *et al.*, 2021) self-sufficient technology, like IoT in case sensors (Shrestha *et al.*, 2022; Rehman *et al.*, 2024). Electrostatic, piezoelectric and electromagnetic solutions are the main techniques for recovering energy from vibrations. Their characteristics vary, affecting their range of applications, implementation constraints and efficiency. Both active (Hoić *et al.*, 2024) and semi-active (Wei & Pang, 2023) seat suspension systems can take advantage of recuperative braking. As stated in (Sun *et al.*, 2018), the PMSM motor, serving as an electromagnetic damper, is mounted to the vehicle body and the unsprung mass. By connecting motor phases in different series-parallel resistor configurations, this solution offers semi-active control of the seat suspension system and enables the control of the suspension damping force through an external circuit with resistors and MOSFET drivers.

This paper expands on the concept of utilizing a brushless direct current (BLDC) motor as a force generator to counteract seat suspension forces or as an energy harvester in the regenerative braking mode. In summary, harnessing supplementary energy from operational devices

Ministry of Science and Higher Education
Republic of Poland

This publication has been funded by the Polish Ministry of Science and Higher Education under the Excellent Science II programme "Support for scientific conferences".
The content of this article was presented during the 61st Symposium "Modelowanie w mechanice" (Modelling in Mechanics), Szczyrk, Poland, March 2–5, 2025.

is a superior alternative to utilizing conventional batteries as a power source for equipment, as it yields a lower overall energy value delivered to the system, thereby enhancing the energy utilization coefficient and reducing operational costs.

2. Model of horizontal seat suspension and hardware implementation

Figure 1a illustrates a physical representation of a horizontal seat suspension system featuring an active control mode and an energy harvesting device. The passive system comprises two tension springs operating in opposing directions, facilitating the establishment of a static equilibrium position for a seat suspension burdened by the suspended mass. A hydraulic damper is utilized to diminish the amplitude of resonant vibrations in the passive system. This system operates effectively under low-friction conditions, facilitated by needle bearings in the suspension mechanism. Additionally, the seat suspension is equipped with end-stop buffers that restrict movement to a maximum displacement of the suspension system. The active system comprises an induction motor for active vibration control (active motoring mode) and for energy harvesting (regenerative braking mode). Figure 1b illustrates the actual experimental apparatus employed to evaluate the vibro-isolation characteristics of the seat. The seat is affixed to a test rig equipped with mechanical components that replicate actual vibrations and forces. Dimensions of the suspension mechanism are as follows: length 425 mm, width 255 mm, height 225 mm. Its unsprung mass, in turn, is about 8 kg and the sprung mass is approximately 7 kg. The configuration comprises sensors and actuators to assess the reaction of the seat to vibrations, guaranteeing that the active suspension system can efficiently lessen undesirable oscillations.

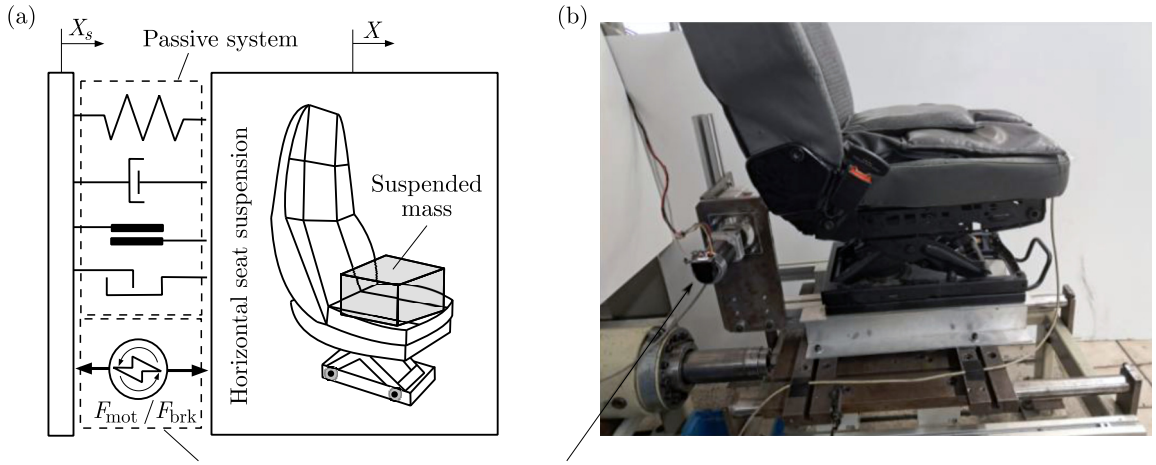

Fig. 1. Physical model of horizontal seat suspension system with active motoring and regenerative braking (a) and set-up for experimental investigation of the vibro-isolation effectiveness (b).

Energy harvesting occurs solely when the intended active force F_{mot} exhibits a sign contrary to the relative velocity $\dot{x} - \dot{x}_s$ of the suspension system (where $\dot{x}$ represents the velocity of seat and $\dot{x}_s$ denotes the velocity of input vibration). In the opposite situation, the vibration reduction system operates by drawing electricity from an external energy source. To generate an active force, a brushless three-phase electric motor cooperates with the system for harvesting energy from mechanical vibrations (Fig. 2). The motor operation is controlled by a dedicated controller that regulates parameters based on input signals, such as torque and rotation direction. The analogue regenerative braking signal, originating from the system controller, is converted into a digital form by a microcontroller. Then, one of five signals generated in this process (ST_1, ST_2, ST_3, ST_4, ST_5) is used to control one of the braking resistor groups (R_1, R_2, R_3, R_4, R_5) via MOSFET transistors.

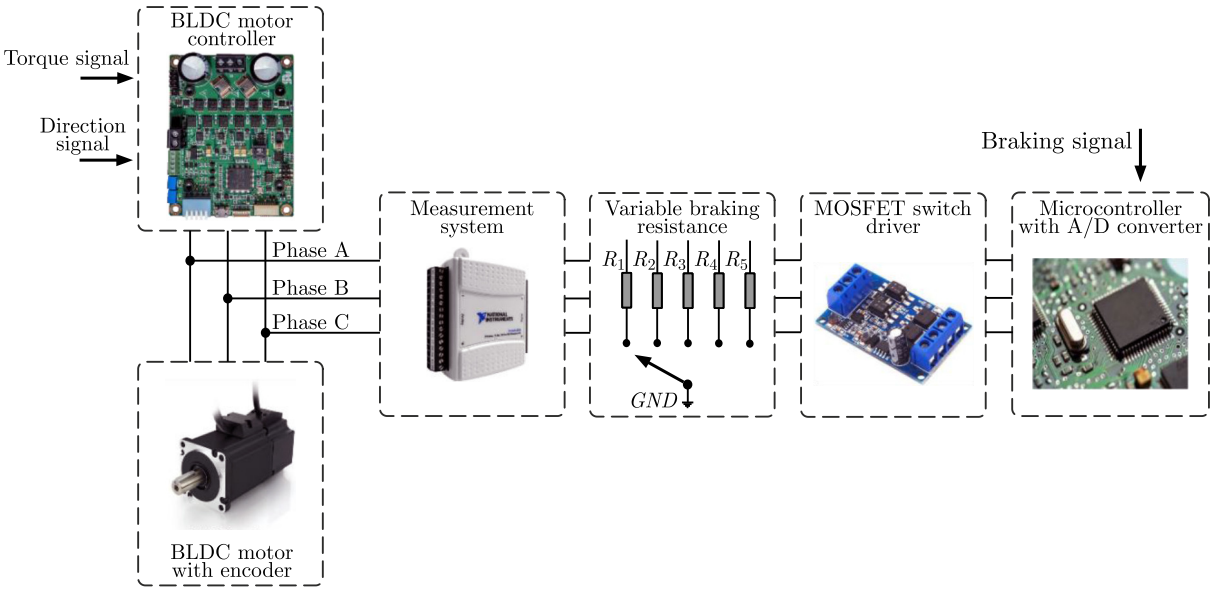

Fig. 2. Hardware implementation of the energy recovery braking subsystem.

The intensity of regenerative braking is regulated through the binary selection of one of five braking resistor groups with resistances of: 0.1 Ω for R_1, 0.15 Ω for R_2, 0.22 Ω for R_3, 0.33 Ω for R_4, and 0.47 Ω for R_5 that was determined experimentally to obtain proportionally varying braking force values (lower resistance is equal to higher braking force). Resistor groups are connected to the phases of the BLDC motor. At any given moment, only one resistor group is activated, allowing control over braking force and achieving various levels of energy recovery. The selection process provides straightforward regulations, although it may limit the smooth adjustment of braking force in real-time. The switching time between resistor groups is a key factor influencing system efficiency, as it determines the responsiveness to changes in the regenerative braking signal. The induced currents and voltages on the BLDC motor braking resistors are recorded by the data acquisition system, enabling their analysis and potential optimization of the energy recovery process.

3. Experimental research

Figure 3 shows white noise excitation signals, which were used to excite the dynamic response of the tested system. Figure 3a presents the time course of the displacement. All three signals exhibit a similar random character, but there are slight differences in amplitude. The waveforms are characterized by maximum deflections in the range of ±0.025 m. Figure 3b shows spectra of the same signals in the frequency domain. Their energy is distributed across a frequency band up to 10 Hz. The differences in power spectral density (PSD) levels indicate variable excitation intensity in individual samples. The research detailed in the paper was conducted using these

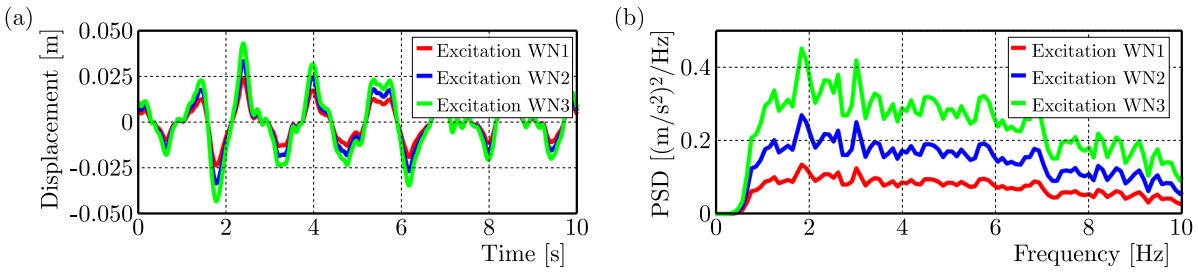

Fig. 3. Displacement of the random input vibration (a) and the PSD of acceleration signal (b) at different excitation intensities (WN1–WN3).

signals, and each test was carried out for a duration of three minutes. The tests are performed by using the mass load up to 80 kg. It directly loads the seat (reproducing upper part of the human body) that reflects the mass of a person weighing approximately 105 kg since the operator supports himself with his limbs.

Figure 4 presents the transmissibility functions of passive (Figs. 4a, 4b), active (Figs. 4c, 4d), and regenerative (Figs. 4e, 4f) seat suspension systems under different excitation intensities (WN1–WN3) and mass loads (40 kg and 80 kg). The transmissibility is plotted against the frequency (0 Hz–10 Hz), which is the key range for ride comfort analysis. The passive seat suspension system in Fig. 4a (WN1 (40 kg) versus WN3 (80 kg)) shows that at low excitation intensity (WN1) with a smaller load (40 kg), the transmissibility is slightly lower at resonance approx. 1 Hz–2 Hz, the red line, compared to the heavier load (80 kg) and higher excitation intensity (WN3), the blue line. Both curves exhibit a clear resonance peak around 1 Hz–2 Hz, typical of passive systems. Transmissibility decreases beyond the resonance frequency in both cases, but faster in the case of WN3 (80 kg). Figure 4b shows the case WN1 (80 kg) versus WN3 (40 kg). The clear peak transmissibility is higher for the smaller load (40 kg) with a higher excitation (WN3), the blue line, than for the heavier load (80 kg) with a lower excitation (WN1), the red line. This is characteristic for very low frequencies (below 1 Hz). Behind this range both curves exhibit a clear resonance peak around 1 Hz–2 Hz, with a slight shift (towards 2 Hz) of the maximum value for the WN3 (40 kg) configuration. This indicates sensitivity of passive systems to mass and excitation changes.

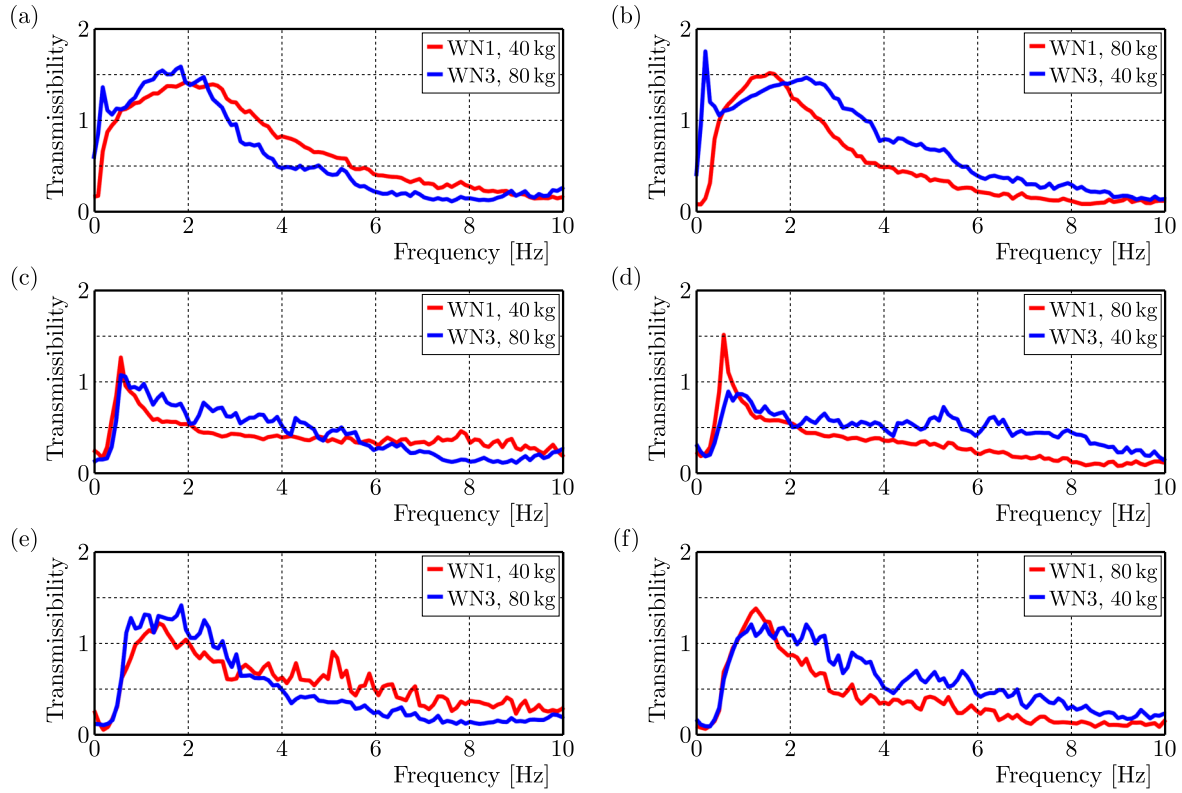

Fig. 4. Transmissibility functions of the passive (a)–(b), active (c)–(d), and regenerative (e)–(f) seat suspension system at different excitation intensity: WN1–WN3, and at various mass loads: 40 kg–80 kg.

In Fig. 4c the active seat suspension system response is presented, where two configurations WN1 (40 kg) versus WN3 (80 kg) are compared. The resonance peak is notably reduced compared to passive systems, showing active damping effectiveness. Overall lower transmissibility, especially at 1 Hz–3 Hz, is observed. The system performs better at lower excitation and lighter mass (red line) but still controls vibration effectively at higher intensity and mass. For the

configuration WN3 (80 kg), the active system seems to perform better above 6 Hz. Figure 4d (WN1 (80 kg) versus WN3 (40 kg)) shows a similar trend, i.e., active control reduces transmissibility across frequencies. The clear peak far above the transmissibility value of 1 occurs for the configuration WN1 (80 kg). Slight variations in performance depend on load and excitation, but overall stability is maintained. The regenerative seat suspension system is shown in Fig. 4e for configurations WN1 (40 kg) versus WN3 (80 kg). It is comparable to the active system in terms of reduced resonance peaks. It maintains lower transmissibility at mid and high frequencies (3 Hz–10 Hz) and shows good performance even at higher mass and excitation. In Fig. 4f (WN1 (80 kg) versus WN3 (40 kg)), resonance damping is slightly less effective than the purely active system but better than the passive one. This shows balance between damping and energy recovery without significant compromise in comfort.

The values in Table 1 provide numerical insights into the regenerative performance (in terms of RMS current and power) of a regenerative seat suspension system under varied excitation intensities (WN1–WN3) and mass loads (40 kg–80 kg). These results correlate directly with the transmissibility trends shown in Figs. 4e, 4f, offering a combined perspective on both ride comfort and energy harvesting potential.

Table 1. Numerical values of the regenerated RMS current and RMS power at different excitation intensity: WN1–WN3, and at various mass loads: 40 kg–80 kg.

Input vibration	Mass load					
	40 kg		60 kg		80 kg	
	RMS current	RMS power	RMS current	RMS power	RMS current	RMS power
WN1	1.020 A	2.130 W	1.152 A	2.428 W	1.195 A	2.613 W
WN2	1.714 A	4.229 W	1.849 A	4.584 W	2.061 A	5.735 W
WN3	2.338 A	6.401 W	2.502 A	7.009 W	2.654 A	7.692 W

In Table 1, the proportional effect of excitation intensity (from WN1 to WN3) is seen in that both RMS current and power increase monotonically with higher excitation, across all mass loads. This is expected as higher excitation introduces more energy, allowing the regenerative system to harvest more. The effect of the mass load (from 40 kg to 80 kg) is noticed as well because RMS current and power increase with heavier loads for each excitation level.

4. Conclusions

This research clearly demonstrates that active and regenerative suspension systems outperform passive systems in maintaining lower transmissibility across frequencies and under different mass and excitation conditions. Overall, heavier mass increases inertial forces, causing greater relative motion in the suspension system and leading to higher energy recovery. The increase is nonlinear but consistent, suggesting the system scales well with input energy and mass. Even under high excitation, the regenerative system maintains low transmissibility, indicating again the balance between comfort and power generation. Subsequent efforts will concentrate on enhancing the ratio of recovered energy to energy consumed during operation. This will be accomplished by enhancing the control algorithm and optimizing the electrical system that manages the transition between power and braking states.

References

1. Hoić, M., Kranjčević, N., & Birt, D. (2024). Design of an active seat suspension based on the Kempe mechanism. *2024 21st International Conference on Mechatronics – Mechatronika (ME)*, 1–7. https://doi.org/10.1109/ME61309.2024.10789764

2. Paul, K., Amann, A., & Roy, S. (2021). Tapered nonlinear vibration energy harvester for powering Internet of Things. *Applied Energy*, *283*, Article 116267. https://doi.org/10.1016/j.apenergy.2020.116267

3. Rehman, S.U., Usman, M., Toor, M.H.Y., & Hussaini, Q.A. (2024). Advancing structural health monitoring: A vibration-based IoT approach for remote real-time systems. *Sensors and Actuators A: Physical*, *365*, Article 114863. https://doi.org/10.1016/j.sna.2023.114863

4. Shrestha, K., Sharma, S., Pradhan, G.B., Bhatta, T., Rana, S.S., Lee, S., Seonu, S., Shin, Y., & Park, J.Y. (2022). A triboelectric driven rectification free self-charging supercapacitor for smart IoT applications. *Nano Energy*, *102*, 107713. https://doi.org/10.1016/j.nanoen.2022.107713

5. Sun, S., Dai, X., Wang, K., Xiang, X., Ding, G., & Zhao, X. (2018). Nonlinear electromagnetic vibration energy harvester with closed magnetic circuit. *IEEE Magnetics Letters*, *9*, 1–4, Article 6102604. https://doi.org/10.1109/LMAG.2018.2822625

6. Wei, C., & Pang, X. (2023). Modeling and simulation for a novel semi-active seat-suspension with a cam-roller-spring mechanism. *2023 6th International Conference on Intelligent Robotics and Control Engineering (IRCE)*, 108–112. https://doi.org/10.1109/IRCE59430.2023.10254784

Manuscript received June 16, 2025; accepted for publication September 8, 2025; published online September 20, 2025.

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 815–822, Warsaw 2025
https://doi.org/10.15632/jtam-pl/210752

TRIBOLOGICAL EFFECTS OF HYALURONIC ACID CONCENTRATION ON ARTICULAR CARTILAGE: PIN-ON-PLATE FRICTION TESTS IN PORCINE AND OSTEOARTHRITIC HUMAN TISSUE

Adam J. MAZURKIEWICZ[1*], Celina PEZOWICZ[2], Anna NIKODEM[2], Maciej PRZYBYŁEK[3], Dominika MAZURKIEWICZ[4]

[1] *Faculty of Mechanical Engineering, Bydgoszcz University of Science and Technology, Bydgoszcz, Poland*
[2] *Faculty of Mechanical Engineering, Wrocław University of Science and Technology, Wrocław, Poland*
[3] *Faculty of Pharmacy, Nicolaus Copernicus University in Toruń, Toruń, Poland*
[4] *Faculty of Medicine, Medical University of Gdańsk, Gdańsk, Poland*

*corresponding author, *adam.mazurkiewicz@pbs.edu.pl*

Hyaluronic acid (HA) is the main biopolymer used in intra-articular injections for osteoarthritis (OA). HA is usually applied at 1–3 mg/mL, though the optimal level remains unclear. The purpose of this study was to determine the effect of HA concentration on the friction in osteoarthrosis.

Samples from the osteoarthritic head of a human femur and porcine controls were tested in a pin-on-plate setup. The results showed a statistically significant effect of HA concentration on the friction in a group of porcine cartilage. In a group of osteoarthritic cartilage, such a relationship did not occur. This comparison highlights that degeneration limits HA's effect.

Keywords: osteoarthrosis; hyaluronic acid; joint friction; tribology; biomechanics.

1. Introduction

Osteoarthritis (OA) is caused by aging joints and is increasingly influenced by lifestyle choices, lack of physical activity, and a diet based on highly processed foods (Allen & Golightly, 2015). It is one of the main causes of functional disability. It can occur in all joints, but the effects of OA of the hand, knee, hip and spine are usually considered the most dangerous. The symptoms of this disease include joint pain, stiffness, and difficulty in movement. This is the result of changes in the structure of the articular cartilage, which loses its physiological shape, structure, and properties. As a result of these changes, pain occurs during movement, resulting in limited mobility for patients (Burr & Gallant, 2012; Aitken *et al.*, 2020). In order to improve mobility in the affected joint, supplementation with hyaluronic acid (HA) preparations is often used to relieve pain and improve its lubricating conditions (Gaumet *et al.*, 2018; Turajane *et al.*, 2007). HA is a component of synovial fluid (SF) responsible for the proper functioning of joints. In a healthy joint, HA has lubricating and shock absorbing functions, reducing friction between surfaces. HA provides the joint lubricant with high viscosity and elasticity, which protects it against mechanical overload. Various authors have analyzed the effect of HA concentration and

Ministry of Science and Higher Education
Republic of Poland

This publication has been funded by the Polish Ministry of Science and Higher Education under the Excellent Science II programme "Support for scientific conferences".
The content of this article was presented during the 61st Symposium "Modelowanie w mechanice" (Modelling in Mechanics), Szczyrk, Poland, March 2–5, 2025.

molar mass on the coefficient of friction (CoF) of articular cartilage. De Roy *et al.* (2024) and Snetkov *et al.* (2020) showed that cartilage friction is mainly determined by its microscopic structure, while viscoelastic properties are additionally related to macroscopic structure. Viscoelastic and frictional properties showed a weak correlation. Caligaris *et al.* (2009) studied the effect of OA degeneration on the friction coefficient value. They assessed friction on seven specimens with a degeneration stage ≤ 2 and nine specimens >2 ≤ 3 on the ICRS scale. They found no statistically significant differences between the friction coefficient value and the degree of OA. Neu *et al.* (2010) investigated that CoF of femoral cartilage samples correlates positively with the severity of OA. These results are not consistent, in addition these studies were conducted on HA solutions produced under laboratory conditions, not on HA preparations used for SF supplementation by injection into the joint.

The aim of the study: In clinical practice, the most commonly used preparations contain between 1 mg/mL and 3 mg/mL of hyaluronic acid. However, from a medical point of view, there are no clear recommendations for injecting a preparation containing a specific concentration of HA (Snetkov *et al.*, 2020; Jin & Dowson, 2013). The study conducted here was designed to answer how the concentration of hyaluronic acid in an HA preparation for injection into the joint affects the reduction of cartilage friction. The lubrication efficacy of preparations with different hyaluronic acid contents was evaluated based on the value of the friction coefficient between the articular cartilage and surgical stainless steel.

2. Material and methods

2.1. Material

The study used 18 osteoarthritic cartilage samples taken from 9 heads of osteoarthritic human femur. Two cylindrical specimens of 10 mm in diameter and 15 mm in height were taken from each head. The femoral heads were obtained from patients undergoing hip replacement.

A single freeze-thaw protocol was applied for sample preservation. The heads were immediately frozen at $-22\,^{\circ}$C after collection. Before examination, they were thawed for 8 hours at $23\,^{\circ}$C, followed by sampling and examination. The storage protocol was selected based on the findings of Szarko *et al.* (2010), who demonstrated that freezing articular cartilage at $-20\,^{\circ}$C or $-80\,^{\circ}$C, followed by controlled thawing at room temperature, maintains the tissue's mechanical properties without causing significant changes.

Figure 1 shows the process of extracting samples for testing. The authors had permission from the local ethics committee to collect and use the material for the study. As a lubricant, a commercially available intra-articular injection product containing 2.2 % high-molecular-weight hyaluronic acid was used. To obtain lower concentrations, the product was diluted with deionized water.

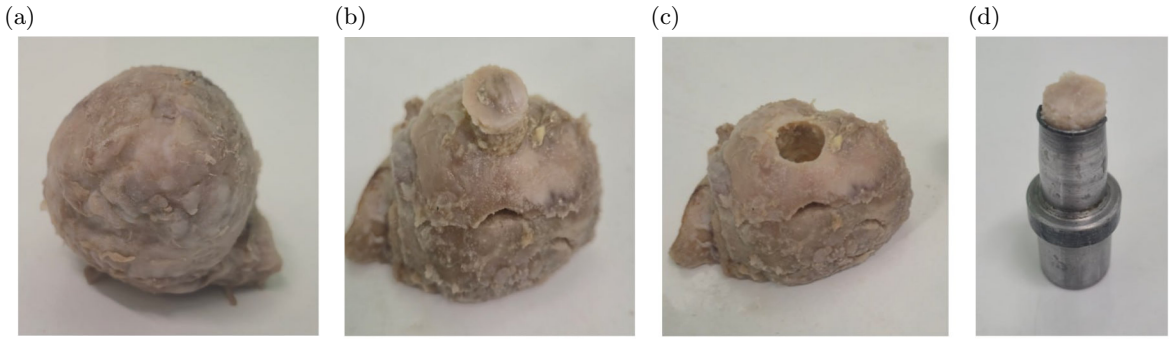

Fig. 1. Steps of sampling for testing: (a) osteoarthritic femoral head; (b), (c) specimen extraction; (d) collected sample.

The control group consisted of 18 samples of pig cartilage taken from pig femoral heads. The bones were obtained from a local slaughterhouse from pigs of the Polish White breed. The heads cut from the bones were frozen immediately after slaughter. The samples were prepared for testing and stored in the same way as human bones.

2.2. Friction coefficient measurements

The tests were conducted using the pin-on-plate method. In this method, an articular cartilage sample was mounted on a stationary pin, while a flat stainless steel counter-sample moved in a rectilinear motion on a moving table. A lubricant containing HA was placed between the samples. Compared with the pin-on-disk method, the principal advantage of the pin-on-plate method is that it ensures a constant relative linear speed between the sample and the counter-sample. In the pin-on-disk method, a stationary pin is in contact with a rotating disk. The disadvantage of this method is that the linear velocity of relative motion between the samples depends on the distance from the axis of rotation of the disk. The linear velocity of the subarea of the sample located on the axis of the rotating disk is zero, whereas subareas farthest from the axis of rotation have the maximum linear velocity. Therefore, the test using the pin-on-plate method more closely reflects the real conditions of movement in the joint. The device used for the study was described by Gordon *et al.* (2014).

ISO 7206 and ISO 14242 series of standards are frequently used standards for evaluating the wear characteristics of hip implants. They specify methods of measurement, values of loads used for testing, directions of application, environmental conditions of testing and others. Based on an analysis of the parameter values recommended in these standards and those used by other researchers (Furmann *et al.*, 2020; Caligaris *et al.*, 2009), a dedicated test program was developed. The speed of movement between the two samples was 0.05 m/s, which corresponds to slow walking (Furmann *et al.*, 2020). Each test was divided into 5 cycles. Each cycle contained 2 steps: movement and rest. The movement time was 2 seconds followed by a 2-second break. Therefore, one test contained 5 cycles of movement and rest. This was to reflect the way the femoral head is loaded during walking, when it is loaded with body weight in the stance phase and unloaded in the swing phase. The reciprocal pressing force of the samples was 10 N (Furmann *et al.*, 2020).

To lubricate the surfaces, preparations used for intra-articular injections containing HA at concentrations of: 1.0 %, 1.5 %, 1.8 %, 2.0 %, and 2.2 % HA were used. The preparations did not contain other substances that can affect the coefficient of friction.

3. Results

Tables 1 and 2 show the mean value of the coefficient of friction, median, standard deviation, minimum and maximum values measured for the sample groups tested (pork cartilage and osteoarthritic cartilage). Additionally, Fig. 2 presents these data as bar charts, which allows a direct visual comparison between the two groups.

Table 1. Friction coefficient values for pork cartilage – stainless steel pairs.

Parameter	HA concentration in lubricant				
	1.0 %	1.5 %	1.8 %	2.0 %	2.2 %
Mean value of the friction coefficient, μ	0.019	0.016	0.015	0.013	0.012
Standard deviation, SD	0.005	0.004	0.003	0.004	0.003
Minimum value	0.012	0.011	0.011	0.005	0.006
Maximum value	0.029	0.023	0.022	0.020	0.019
Median	0.019	0.016	0.015	0.013	0.0115

Table 2. Friction coefficient values for osteoarthritic cartilage – stainless steel pairs.

Parameter	HA concentration in lubricant				
	1.0 %	1.5 %	1.8 %	2.0 %	2.2 %
Mean value of the friction coefficient, μ	0.033	0.031	0.029	0.028	0.026
Standard deviation, SD	0.007	0.007	0.004	0.004	0.007
Minimum value	0.021	0.021	0.024	0.021	0.019
Maximum value	0.044	0.042	0.037	0.035	0.040
Median	0.033	0.032	0.031	0.029	0.025

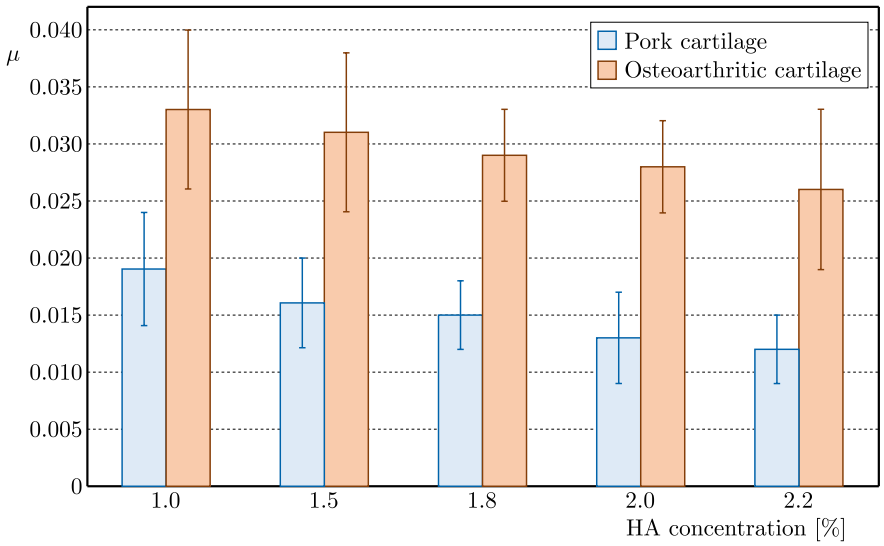

Fig. 2. Effect of hyaluronic acid concentration on the friction coefficient (μ) in porcine and osteoarthritic human cartilage (means with SD error bars).

A statistical analysis of the results was carried out to assess the differences in friction coefficients for lubricants with different HA contents. As a first step, the Kolgomorov–Smirnov and Levene's test was performed at a significance level of $\alpha = 0.05$ to check the type of distribution of results. In each sample group, the results had a normal distribution and equal variances. Further analyses of the significance of differences in the coefficient of friction for lubricants with different HA contents were performed using Anova's one-way analysis at a significance level of $\alpha = 0.05$. Tukey's test was used to determine which groups had statistically significant differences in the mean values of the friction coefficient. The results of the statistical tests are shown in Tables 3 and 4. All analyses were performed using Statistica 13 software (StatSoft, PL).

Table 3. Tukey's test results for the pork cartilage – stainless steel pair.

HA concentration	1.0 %	1.5 %	1.8 %	2.0 %	2.2 %
1.0 %	–	NS	S	S	S
1.5 %	–	–	NS	NS	S
1.8 %	–	–	–	NS	S
2.0 %	–	–	–	–	NS
2.2 %	–	–	–	–	–

An S value means that the difference in mean values between the two groups is statistically significant. An NS value means that statistically, there is no difference between the mean values in two particular groups.

Table 4. Tukey's test results for the osteoarthritic cartilage – stainless steel pair.

HA concentration	1.0 %	1.5 %	1.8 %	2.0 %	2.2 %
1.0 %	–	NS	NS	NS	NS
1.5 %	–	–	NS	NS	NS
1.8 %	–	–	–	NS	NS
2.0 %	–	–	–	–	NS
2.2 %	–	–	–	–	–

4. Discussion

The friction coefficient values obtained from the study for the pork cartilage-steel friction pair were in the range of 0.019–0.012, while for human osteoarthritic cartilage, they were in the range of 0.033–0.026. For pork cartilage, these values are in line with literature data, i.e., 0.005–0.02 (Furmann *et al.*, 2020; Jin & Dowson, 2013). For osteoarthritic cartilage, these values are higher than for healthy cartilage but comparable with results obtained by other authors (Caligaris *et al.*, 2009).

In both test groups, the average value of the coefficient of friction decreased as the concentration of HA in the lubricant increased. However, statistical analyses showed that the effect of HA concentration in the lubricant on the coefficient of friction was significant only for the pork cartilage. No such relationship was found in the OA group. Porcine samples were the reference group, as cartilage in samples from this group had no pathological changes. The use of human articular cartilage without pathological features in the reference group was not possible due to the lack of approval from the local ethics committee. Nevertheless, porcine cartilage is widely accepted as a suitable model for human articular cartilage (Fackler *et al.*, 2023). Furthermore, the mechanical properties of swine cartilage, including stiffness under defined loading conditions, have been reported to approximate those of human tissue, further supporting its application in biomechanical evaluations (Ronken *et al.*, 2012).

The statistically significant differences, or lack thereof, observed in the study can be explained at the molecular level. At this scale, the variation in concentration-dependent response is primarily determined by the condition of the cartilage surface. In healthy tissue, densely hydrated hyaluronic acid coils adsorb onto the phospholipid-rich superficial zone, forming electrostatic interactions with both phosphatidylcholine head groups and the underlying collagen network. This promotes the formation of a continuous hydration film that substantially reduces friction. In contrast, degenerative changes associated with OA disrupt this lipid–protein interface and expose denatured collagen fibrils, thereby limiting the availability of effective HA-binding sites, which likely accounts for the absence of statistically significant differences observed in osteoarthritic tissue. This interpretation is supported by findings from NMR-based compression experiments, which showed that enzymatic degradation of the collagen fibrillar network leads to mechanical softening and an almost complete loss of swelling capacity due to impaired fluid pressurization and a disrupted pore structure (Greene *et al.*, 2012). These structural alterations reduce the tissue's ability to interact effectively with HA and to maintain a functional lubrication environment under load.

It is important to emphasize that certain methodological challenges are inherent to studies involving biological tissues, such as cartilage. In the case of cartilage, as in the case of the study of other tissues (Kohut *et al.*, 2021; Aleksandrowicz, 2020), the evaluation of biomechanical characteristics by methods used to test structural materials is not straightforward, and the accuracy of measurement may be unsatisfactory. This is due to the specific characteristics of the material, the difficulty of determining the actual way in which the tissue is loaded in the body, and choosing the correct method of conducting the test.

Cartilage lubrication in the joint occurs in two ways: by compression of the interstitial fluid (Ateshian *et al.*, 1998; Krishnan *et al.*, 2004) and boundary lubrication by the SF (Schmidt *et al.*, 2007a; Schmidt & Sah, 2007b). Caligaris *et al.* (2009) showed that lubrication by compression of interstitial fluid is usually much more effective than boundary lubrication by SF. During OA, the structure of collagen fibers in the upper layers of the particle is damaged, and consequently, its porosity and permeability are higher. Therefore, during cartilage deformation, the increase in fluid pressure in the cartilage matrix with OA is not as great as in healthy cartilage.

In our study, the cartilage samples were 10 mm in diameter, while the steel counter-sample was flat and smooth. Due to the spherical structure of the articular surfaces, it is impossible to obtain relatively flat specimens with larger dimensions on which to perform a more accurate test. The dimensions of the specimen made it difficult to obtain the correct fluid pressure in the cartilage, due to the extrusion of fluid from the specimen and the lack of fluid flow throughout the cartilage. This could also have affected the accuracy of the measurement.

The next factor to analyze was the speed at which the test was conducted. Tests were conducted at the speed of reciprocal surface motion corresponding to slow walking, i.e., 0.05 m/s. At other speeds, due to the non-Newtonian nature of the fluid, the friction coefficient values may be different.

Another factor is changes in morphology in the subchondral layer and the trabecular bone that supports the cartilage. As a result of OA, the shape, structure, as well as mineral content of these tissues may change (Cichański *et al.*, 2010; Topoliński *et al.*, 2012a; 2012b). As a result, the elasticity of the cartilage may also change.

Balazs (2004) showed that the intramedullary injection of HA improves the viscoelasticity and fluidity of SF, alleviates the effects of OA, prevents symptoms of the disease, and allows postponement of surgery. However, it is difficult to determine whether the concentration of injectable HA affects the duration of effective impact. It is highly dependent on the individual characteristics of the patient and many factors, such as the degree of joint damage, the patient's weight, and level of physical activity.

5. Conclusions

Frictional performance of articular cartilage reflects the interplay between tissue condition and the properties of the lubricating medium. To provide a clear, application-oriented summary, we evaluated how stepwise changes in HA concentration affect the coefficient of friction using a standardized pin-on-plate protocol within a range relevant to viscosupplementation practice.

In pin-on-plate friction testing, increasing hyaluronic-acid concentration from 1.0 % to 2.2 % was associated with a progressive reduction of the friction coefficient in porcine articular cartilage, with several pairwise comparisons reaching statistical significance. In osteoarthritic human cartilage, friction remained consistently higher across the same concentration range and between-concentration differences did not reach significance under the present protocol.

These findings, obtained within a concentration range commonly used in viscosupplementation, highlight the practical importance of reporting and controlling HA content in tribological assessments. For non-degenerate tissue, higher HA levels can yield a tangible reduction in friction; for osteoarthritic tissue, adjusting HA concentration alone may be insufficient, suggesting the value of exploring more physiologically representative lubricants or combined approaches. Future work should expand the number of specimens per concentration and examine broader loading and speed conditions.

Acknowledgments

The research described in this article was carried out as part of a program of research internship completed by the first author of the paper (A. Mazurkiewicz) at the Faculty of Mechanical Engineering of Wroclaw University of Science and Technology (Poland).

References

1. Aitken, D., Jones, G., & Winzenberg, T.M. (2020). Clinical overview of osteoarthritis (OA) and the challenges faced for future management. In S.I.S. Rattan (Ed.), *Encyclopedia of Biomedical Gerontology* (pp. 420–430). Academic Press. https://doi.org/10.1016/B978-0-12-801238-3.11419-9
2. Aleksandrowicz, P. (2020). Modeling head-on collisions: The problem of identifying collision parameters. *Applied Sciences*, *10*(18), Article 6212. https://doi.org/10.3390/app10186212
3. Allen, K.D., & Golightly, Y.M. (2015). State of the evidence. *Current Opinion in Rheumatology*, *27*(3), 276–283. https://doi.org/10.1097/BOR.0000000000000161
4. Ateshian, G.A., Wang, H., & Lai, W.M. (1998). The role of interstitial fluid pressurization and surface porosities on the boundary friction of articular cartilage. *Journal of Tribology*, *120*(2), 241–248. https://doi.org/10.1115/1.2834416
5. Balazs, E.A. (2004). Viscosupplementation for treatment of osteoarthritis: from initial discovery to current status and results. *Surgical Technology International*, *12*, 278–289.
6. Burr, D.B., & Gallant, M.A. (2012). Bone remodelling in osteoarthritis. *Nature Reviews Rheumatology*, *8*, 665–673. https://doi.org/10.1038/nrrheum.2012.130
7. Caligaris, M., Canal, C.E., Ahmad, C.S., Gardner, T.R., & Ateshian, G.A. (2009). Investigation of the frictional response of osteoarthritic human tibiofemoral joints and the potential beneficial tribological effect of healthy synovial fluid. *Osteoarthritis and Cartilage*, *17*(10), 1327–1332. https://doi.org/10.1016/j.joca.2009.03.020
8. Cichański, A., Nowicki, K., Mazurkiewicz, A., & Topoliński, T. (2010). Investigation of statistical relationships between quantities describing bone architecture, its fractal dimensions and mechanical properties. *Acta of Bioengineering and Biomechanics*, *12*(4), 69–77.
9. Comper, W.D., & Laurent, T.C. (1978). Physiological function of connective tissue polysaccharides. *Physiological Reviews*, *58*(1), 255–315. https://doi.org/10.1152/physrev.1978.58.1.255
10. De Roy, L., Teixeira, G.Q., Schwer, J., Sukopp, M., Faschingbauer, M., Ignatius, A., & Seitz, A.M. (2024). Structure-function of cartilage in osteoarthritis: An ex-vivo correlation analysis between its structural, viscoelastic and frictional properties. *Acta Biomaterialia*, *190*, 293–302. https://doi.org/10.1016/j.actbio.2024.10.027
11. Fackler, N.P., Donahue, R.P., Bielajew, B.J., Amirhekmat, A., Hu, J.C., Athanasiou, K.A., & Wang, D. (2023). Characterization of the age-related differences in porcine acetabulum and femoral head articular cartilage. *Cartilage*, *16*(3), 366–375. https://doi.org/10.1177/19476035231214724
12. Furmann, D., Nečas, D., Rebenda, D., Čipek, P., Vrbka, M., Křupka, I., & Hartl, M. (2020). The effect of synovial fluid composition, speed and load on frictional behaviour of articular cartilage. *Materials*, *13*(6), Article 1334. https://doi.org/10.3390/ma13061334
13. Gaumet, M., Badoud, I., & Ammann, P. (2018). Effect of hyaluronic acid-based viscosupplementation on cartilage material properties. *Osteoarthritis and Cartilage*, *26*(Supplement 1), S136. https://doi.org/10.1016/j.joca.2018.02.294
14. Gordon, M., Słomion, M., & Mazurkiewicz, A. (2014). Design device for friction coefficient examination of articular cartilage. In *31st Danubia Adria Symposium on Advances in Experimental Mechanics. September 24–27, 2014 – Kempten (Germany). Proceedings* (pp. 90–91). VDI Verein Deutscher Ingenieure. https://owncloud.tuwien.ac.at/index.php/s/2IakbkmqJceQx44?dir=/&editing=false&openfile=true
15. Greene, G.W., Zappone, B., Banquy, X., Lee, D.W., Söderman, O., Topgaard, D., & Israelachvili, J.N. (2012). Hyaluronic acid–collagen network interactions during the dynamic compression and recovery of cartilage. *Soft Matter*, *8*(38), 9906–9914. https://doi.org/10.1039/c2sm26330k
16. Jin, Z., & Dowson, D. (2013). Bio-friction. *Friction*, *1*(2), 100–113. https://doi.org/10.1007/s40544-013-0004-4
17. Kohut, P., Holak, K., Ekiert, M., Młyniec, A., Tomaszewski, K.A., & Uhl, T. (2021). The application of digital image correlation to investigate the heterogeneity of Achilles tendon deformation and

determine its material parameters. *Journal of Theoretical and Applied Mechanics*, *59*(1), 43–52. https://doi.org/10.15632/jtam-pl/127903

18. Krishnan, R., Kopacz, M., & Ateshian, G.A. (2004). Experimental verification of the role of interstitial fluid pressurization in cartilage lubrication. *Journal of Orthopaedic Research*, *22*(3), 565–570. https://doi.org/10.1016/j.orthres.2003.07.002
19. Neu, C.P., Reddi, A.H., Komvopoulos, K., Schmid, T.M., & Di Cesare, P.E. (2010). Increased friction coefficient and superficial zone protein expression in patients with advanced osteoarthritis. *Arthritis & Rheumatology*, *62*(9), 2680–2687. https://doi.org/10.1002/art.27577
20. Ronken, S., Arnold, M.P., Ardura García, H., Jeger, A., Daniels, A.U., & Wirz, D. (2012). A comparison of healthy human and swine articular cartilage dynamic indentation mechanics. *Biomechanics and Modeling in Mechanobiology*, *11*(5), 631–639. https://doi.org/10.1007/s10237-011-0338-7
21. Schmidt, T.A., Gastelum, N.S., Nguyen, Q.T., Schumacher, B.L., & Sah, R.L. (2007a). Boundary lubrication of articular cartilage: Role of synovial fluid constituents. *Arthritis & Rheumatology*, *56*(3), 882–891. https://doi.org/10.1002/art.22446
22. Schmidt, T.A., & Sah, R.L. (2007b). Effect of synovial fluid on boundary lubrication of articular cartilage. *Osteoarthritis and Cartilage*, *15*(1), 35–47. https://doi.org/10.1016/j.joca.2006.06.005
23. Snetkov, P., Zakharova, K., Morozkina, S., Olekhnovich, R., & Uspenskaya, M. (2020). Hyaluronic acid: The influence of molecular weight on structural, physical, physico-chemical, and degradable properties of biopolymer. *Polymers*, *12*(8), Article 1800. https://doi.org/10.3390/polym12081800
24. Szarko, M., Muldrew, K., & Bertram, J.E.A. (2010). Freeze-thaw treatment effects on the dynamic mechanical properties of articular cartilage. *BMC Musculoskeletal Disorders*, *11*, Article 231. https://doi.org/10.1186/1471-2474-11-231
25. Topoliński, T., Cichański, A., Mazurkiewicz, A., & Nowicki, K. (2012a). Applying a stepwise load for calculation of the S-N curve for trabecular bone based on the linear hypothesis for fatigue damage accumulation. *Materials Science Forum*, *726*, 39–42. https://doi.org/10.4028/www.scientific.net/MSF.726.39
26. Topoliński, T., Cichański, A., Mazurkiewicz, A., & Nowicki, K. (2012b). The relationship between trabecular bone structure modeling methods and the elastic modulus as calculated by FEM. *The Scientific World Journal*, *2012*(1), Article 827196. https://doi.org/10.1100/2012/827196
27. Turajane, T., Tanavaree, A., Labpiboonpong, V., & Maungsiri, S. (2007). Outcomes of intra-articular injection of sodium hyaluronate for the treatment of osteoarthritis of the knee. *Journal of the Medical Association of Thailand*, *90*(9), 1845–1852. http://www.jmatonline.com/PDF/90-PB-1845-1852.pdf

Manuscript received July 5, 2025; accepted for publication September 15, 2025; published online October 13, 2025.

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 823–837, Warsaw 2025
https://doi.org/10.15632/jtam-pl/203691

INVESTIGATION OF FLOW CONTROL ON A VERTICAL AXIS WIND TURBINE USING A BIONIC FLAP

Mingtong ZHOU[1], Junwei ZHONG[1*], Qiqi ZHANG[1], Yufeng GAN[1], Chaolei ZHANG[1], Huizhong LIU[2]

[1] *School of Mechanical and Electrical Engineering, Jiangxi University of Science and Technology, Ganzhou, China*
[2] *Jiangxi Province Engineering Research Center for Mechanical and Electrical of Mining and Metallurgy, Jiangxi University of Science and Technology, Ganzhou, China*

[*]corresponding author, *jwzhong0@jxust.edu.cn*

Inspired by the slight lift of bird feathers at the trailing edge under specific conditions, an adjustable bionic flap (BF) was added to a vertical-axis wind turbine (VAWT) to improve its aerodynamic performance. Numerical simulations using the SST turbulence model were conducted to examine the BF's flow control mechanism and how its geometry affects the VAWT's power coefficient. The results show that the BF can evidently improve the power coefficients of the VAWT. Compared with the original VAWT, the power coefficient of the VAWT with an adjustable BF is increased by 45.2 % at $\lambda = 1.75$.

Keywords: vertical axis wind turbine; bionic flap; aerodynamic performance; flow control.

1. Introduction

In recent years, the occurrence of severe weather and climate phenomena has been escalating globally, predominantly attributed to the rise in global temperatures. A primary contributor to this global occurrence is the reliance on mineral fuels, which emit significant amounts of greenhouse gases. Transitioning away from mineral fuels presents one of the most significant challenges of the 21st century. Consequently, wind power has garnered significant interest as a prospective substitute (Rehman *et al.*, 2023; McKenna *et al.*, 2025).

Wind turbines are the primary apparatus used to harness wind energy. They can be divided into two principal groups based on the orientation of their rotating shafts: horizontal-axis wind turbines (HAWTs) and vertical-axis wind turbines (VAWTs). VAWTs present significant advantages compared to HAWTs, including omnidirectional operation, improved structural scalability, and enhanced system stability (Abdolahifar & Zanj, 2025). Owing to these advantages, VAWTs are increasingly favored in urban, remote, and offshore settings. Nevertheless, VAWTs are at present defined by their lower energy conversion efficiency in comparison with HAWTs. Several attempts have been made to enhance the aerodynamic performance of VAWTs through various flow control techniques (Zhao *et al.*, 2022; Tayebi & Torabi, 2024). Many researchers highlight the significance of applying these control methods near the blade's leading edge to impact the onset of flow separation. The vortex generator (VG) is a simple apparatus made up of several mini plates, usually mounted on the suction surface of a blade airfoil near the leading edge. While the mechanisms of VGs on airfoils and HAWTs have been extensively studied, relatively less research has focused on their application in VAWTs. Yan *et al.* (2019) proposed the use of micro VGs with heights less than half of the boundary layer thickness for VAWTs. Flow separation on a VAWT blade tends to happen alternatively on the suction and pressure surfaces. A VG with excessive height can generate additional drag, which may undermine its aerodynamic advantages. Zhong *et al.* (2019) proposed an innovative approach by replacing con-

ventional VGs with an elevated rod positioned in front of the blade's leading edge. Ullah *et al.* (2020) used leading-edge slats to reduce the dynamic stall in urban VAWTs at low wind speeds.

The leading-edge protuberance (LEP) represents another promising flow control technology for VAWTs, inspired by the unique design of humpback whale flippers. The effectiveness of LEPs in suppressing flow separation significantly depends on various geometric parameters, including wavelength and wave amplitude. Inadequate geometric parameters can adversely affect the flow performance of a VAWT. Yan *et al.* (2021) discovered that the wave amplitude plays a more critical role than the wavelength in enhancing aerodynamic performance. Furthermore, Chang *et al.* (2024) examined the impact of spacing between protuberances on the performance of biomimetic VAWT blades and reported that blades with long-wavelength LEPs outperform those with short-wavelength counterparts. Consequently, the geometric parameters of LEPs should be thoroughly designed for optimal application on VAWT blades. A crafted blade equipped with leading-edge protuberances can substantially boost the power generation of a VAWT at low tip-speed ratios (TSRs) (Sridhar *et al.*, 2022). Supplementing the already mentioned passive flow control techniques, a range of active control measures are implemented near the leading edge of VAWT blades to reduce flow separation. Rezaeiha *et al.* (2019) implemented leading-edge slot suction to prevent the bursting and formation of laminar separation bubbles, thereby avoiding the development of dynamic stall vortices and trailing-edge roll-up vortices. However, active control technologies necessitate energy consumption, which requires a careful assessment of efficiency and energy expenditure. Abbasi and Daraee (2024) investigated the combined effects of the installation position and actuator activation timing on flow control in a VAWT and proposed a novel method that involves operating plasma actuators for each blade individually to suppress flow separation while minimizing energy consumption.

Trailing-edge control techniques have demonstrated their effectiveness in suppressing flow separation and improving the power coefficient of VAWTs. Among the most commonly employed trailing-edge control methods are trailing-edge jets (Sun & Huang, 2023), Gurney flaps (GFs) (Chen *et al.*, 2020; Zhu *et al.*, 2021; Liu *et al.*, 2022; Syawitri *et al.*, 2024), and trailing-edge flaps (Ertem *et al.*, 2016; Attie *et al.*, 2022; Han *et al.*, 2023). The GF is a small tab attached to the pressure surface of the blade, which increases the blade's chamber. This configuration generates a pair of counter-rotating vortices downstream of the GF, leading to a negative pressure distribution on the suction surface and a positive one on the pressure surface. Zhu *et al.* (2021) conducted a numerical study examining how the geometric parameters of the GF influence the performance of straight-bladed VAWTs. Chen *et al.* (2020) reported that an active GF yields better performance than a fixed GF. Liu *et al.* (2022) explored the combined effects of the GF and cavity on the aerodynamic efficiency of a straight-bladed VAWT. Syawitri *et al.* (2024) proposed a slit-modified GF aimed at reducing drag in lift-type VAWTs. This GF with a slit created small-scale vortices that quickly broke down large coherent flow structures in the near-wake, thus enhancing the lift-to-drag ratio and bettering the torque production. The trailing-edge flap, which is typically separated from the blade by a slot, is more complex than the GF (Ertem *et al.*, 2016). The high-pressure flow that passes through the slot helps delay flow separation and reduces vortex shedding. Attie *et al.* (2022) studied the impact of critical geometric parameters on VAWT performance via the design of experimental methodologies. Furthermore, Han *et al.* (2023) discussed the synergistic control of pitch and trailing-edge flaps in VAWTs, demonstrating that the coordinated motion of pitch and the flap effectively suppress flow separation and minimize load fluctuations in the turbine.

Inspired by the phenomenon where birds slightly increase their feathers at the trailing edge, we propose an adjustable bionic flap (BF) to manage flow separation on the VAWT blade. The configuration of the blade equipped with the BF is depicted in Fig. 1. The BF can significantly mitigate flow separation on a static blade, resulting in an increased lift-to-drag ratio and reduced flow separation (Ma *et al.*, 2022). However, studies focusing on how a BF influences the aerodynamic characteristics of a VAWT are scarce, and the flow dynamics involved with a rotating

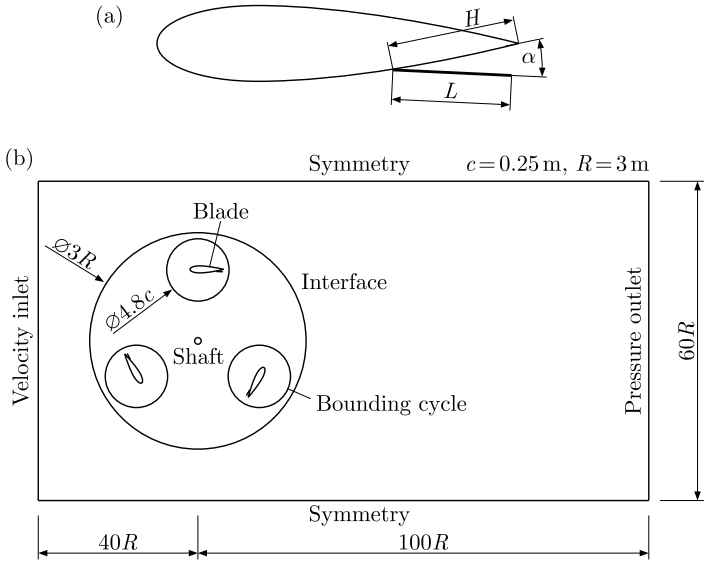

Fig. 1. Geometry models: (a) blade with a BF; (b) computational zone and boundary condition.

VAWT blade equipped with a BF are poorly comprehended. This study utilizes comprehensive numerical simulations conducted with ANSYS Fluent software to explore the flow control mechanisms of the BF on the VAWT blade and evaluate how the geometric parameters of the BF influence the aerodynamic performance of the VAWT. The composition of the paper is set out in the following manner: Section 2 establishes and validates the numerical simulation model; Section 3 discusses the flow control mechanisms of the BF across two typical TSRs; Section 4 investigates the impact of the BF on the power coefficient of the VAWT and the instantaneous torque of a single blade; Section 5 presents the conclusions.

2. Numerical model and validation

2.1. Research subjects

The study focuses on a 12 kW VAWT developed by Uppsala University (Kjellin *et al.*, 2011). This H-type Darrieus wind turbine features a diameter of $R = 3$ m and a height of $H = 5$ m. The blade ends are tapered, starting 1 m from the tip, resulting in a chord length at the tip that is 60 % of that at the midpoint of the blade. The blade airfoil is the NACA 0021, and the chord length at the midpoint is $c = 0.25$ m. The power coefficient (c_p) as a function of the TSR was measured through field testing at two fixed rotational speeds of 48 and 57. The wind shear and the wind distribution of the test site were measured for several years. The c_p-TSR curve was drawn using around 350 h data from a measurement campaign in 3 months. The wind speed range is 0 m/s–11 m/s, which corresponds to a TSR range of 1.75–4.5. This wind turbine was chosen as such a TSR range covers the main operating range of a VAWT, such as deep dynamic stall and light dynamic stall categories. The primary geometric parameters of the BF are illustrated in Fig. 1. The length between the BF hinge point and the blade trailing edge is denoted as H, the pop-up angle is indicated by α, and the BF length is specified as L. The BF was installed on the inner side of the blade as it performs better than that on the outer side. For convenience, the prototype of the 12 kW VAWT is denoted as the original VAWT, and the VAWT controlled by the BF is denoted as the VAWT with the BF.

2.2. Numerical model

Considering the time-intensive nature of 3D unsteady studies, a 2D unsteady numerical model was constructed on the basis of the blade midsection. The computational domain is depicted in

Fig. 1b. This domain spans $40R$ upstream and $100R$ downstream of the VAWT's centre, having a width of $60R$. A velocity inlet boundary condition is assigned to the upstream boundary. A pressure outlet boundary condition is used for the downstream boundary. A symmetry boundary condition is applied to the top and bottom sides, and a no-slip boundary condition is enforced on the blade surface. The computational domain is partitioned into two regions by a circle that has a diameter of $3R$. The internal subdomain is set up as a rotating zone with a rotation speed of 57 rpm, while the external subdomain is designated as a non-rotating zone. A sliding mesh model is implemented at the interfaces between the two subdomains, utilizing a non-conformal interface. Three bounding cycles are established around the three blades to regulate the grid density, with the diameter of the bounding cycles set at $4.8c$.

As depicted in Fig. 2, a hybrid mesh approach is employed to discretize the computational domain via Gambit 2.4.6. A structured quadrilateral grid is utilized for the external subdomain and the boundary layer of the blade, whereas an unstructured triangular grid is applied to the other regions. The BF disrupts the topology around the blade, requiring a subdomain to be split to include the BF. An unstructured triangular grid is used to discretize this particular subdomain, complemented by 20 layers of quadrilateral grid generated along the BF surface. The height of the first layer along both the blade and BF surfaces is set to 1.2×10^{-5} m, ensuring that the wall y^+ is approximately 1 (Rogowski *et al.*, 2018). The unsteady flows around the VAWTs are analyzed via Fluent 16.1 and the SST k-ω model. The SIMPLE algorithm (Patankar & Spalding, 1972) and a second-order upwind scheme are adopted. The residuals of the unsteady calculations are set 10^{-6}. Thirty revolutions were simulated for each case in the grid independence study. The unsteady flow shows periodicity after the fifteenth revolution. Then, the converged flow field was taken as the initial flow field for the cases with different time steps and TSRs and at least ten more revolutions were performed to obtain periodicity.

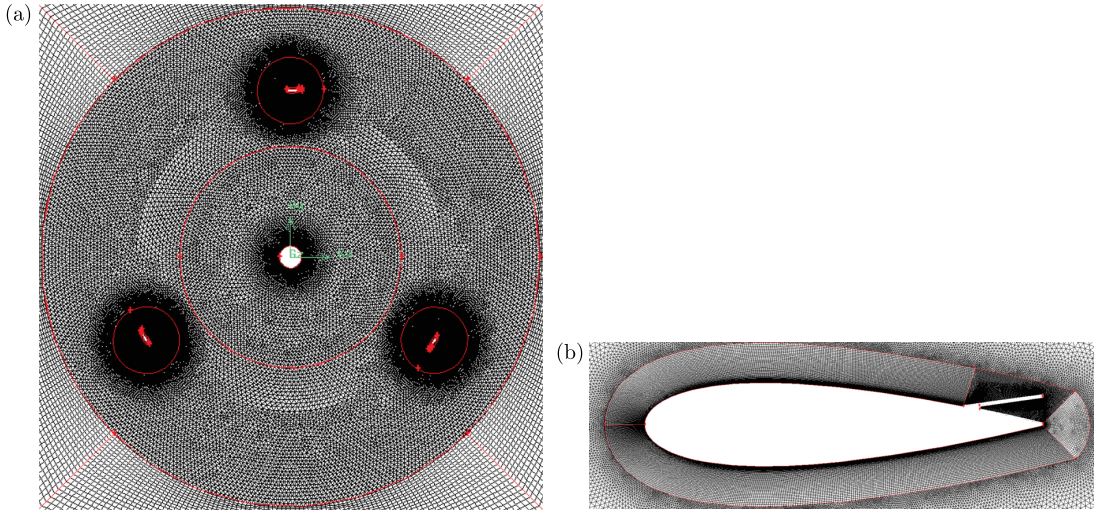

Fig. 2. Grid strategy: (a) grid around the VAWT; (b) gird around the blade with the BF.

The grid independence study was performed on the original VAWT at a TSR of $\lambda = 2.11$, during which the blades undergo deep dynamic stall. The time step was established at $\Delta\theta = 1°$, as the optimal time step is initially unknown. Grid refinement was performed by adjusting the number of cells on the blades, bounding cycles, interfaces, and wakes, as outlined in Table 1. The calculated torques of the VAWT are 54.7 Nm, 56.1 Nm, 59.6 Nm, and 60.7 Nm for the four grid configurations. The relative deviations between Grids 1, 2, 3 and Grid 4 were calculated. The relative deviation decreased with the increase in the grid number. The relative deviation between Grid 3 and Grid 4 dropped to 1.8% and the torque curves of the two grids almost overlapped, as shown in Fig. 3a. Thus, Grid 3 was chosen for the following studies.

Table 1. Grid independence study.

Parameters	Grid 1	Grid 2	Grid 3	Grid 4
Leading-edge spacing [mm]	0.5	0.25	0.1875	0.125
Trailing-edge spacing [mm]	1	0.5	0.375	0.25
Cells on the blade	131	262	370	524
Cells on the bounding cycle	60	180	240	300
Cells on the interfaces	180	360	360	720
Cells on the wake	60	120	180	240
Total number of cells	1.12×10^5	2.73×10^5	3.73×10^5	7.09×10^5
Torque of the VAWT [Nm]	54.7	56.1	59.6	60.7
Relative deviation [%]	−9.88	−7.58	−1.81	–

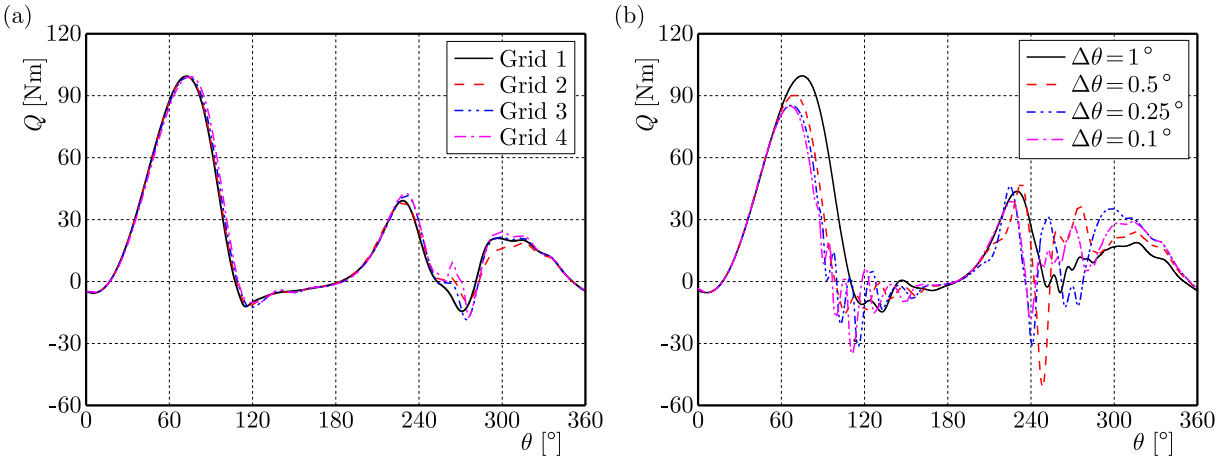

Fig. 3. Results of grid and time independence studies:
(a) grid independence study; (b) time independence study.

A time independence study was conducted on Grid 3 at $\lambda = 2.11$. Four time steps were considered: $\Delta\theta = 1°$, $0.5°$, $0.25°$, and $0.1°$. The torque of a single blade versus the azimuthal angle, θ, is compared in Fig. 3. The torque curve appears relatively smooth when $\Delta\theta = 1°$. This time step is not sufficiently small to capture the fluctuations in torque caused by the shedding of the dynamic stall vortex (Rezaeiha *et al.*, 2018). When the time step is reduced to $\Delta\theta = 0.25°$, fluctuations in the torque are accurately captured. The torque curve subsequently exhibits minimal change as the time step decreases further. The torques of the VAWT at the four time steps are 59.6, 50.7, 48.1, and 48.0 Nm, respectively. Therefore, a time step of $\Delta\theta = 0.25°$ was selected for simulating the flow around the VAWT.

2.3. Model accuracy evaluation

The numerical model was validated by comparing the predicted power coefficient against the experimental ones from Uppsala University (Kjellin *et al.*, 2011), as depicted in Fig. 4. The results are in close agreement with the experiment data at low TSRs. However, a deviation is noted at high TSRs, which is consistent with findings in other studies (Daróczy *et al.*, 2015). These deviations primarily stem from two factors: first, the numerical error brought by discretization schemes, which is inevitable in the process of converting partial differential equations into linear equations; second, the adoption of a 2D numerical simulation, which overlooks losses induced by 3D effects, such as blade tip loss and support arm loss. These 3D losses tend to increase with increasing TSR. Daróczy *et al.* (2015) proposed a correction for these deviations represented by $\Delta C_p^{\text{corr}} = -0.0021\lambda^3$. As shown in Fig. 4, the modified values correspond closely with the

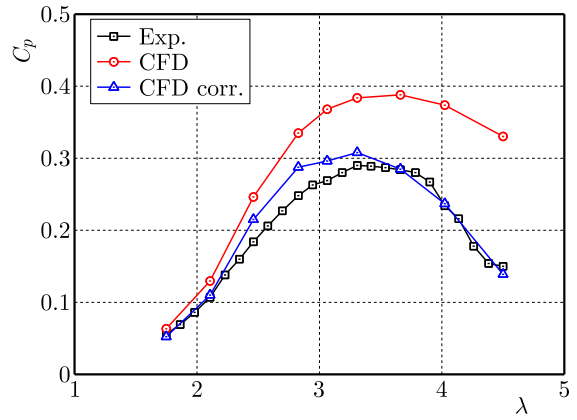

Fig. 4. Validation of the numerical model by power coefficient.

experiment data. The aim of this research is to explore the mechanism of the BF in the flow separation. The influence of the BF on the three-dimensional effects is considered relatively minor and is thus excluded from further discussion. Consequently, the numerical model is deemed sufficiently accurate for the current investigation.

3. Flow control mechanism of the BF

Due to the drag induced by the BF, not well-deigned BF may lead to a deterioration of the VAWT power coefficient for $\lambda > 3.66$. The BF with $L20H20\alpha15$ is chosen to analyze its effect on flow separation, as the case presents a positive effect on the power coefficient of the VAWT from $\lambda = 1.75$ to $\lambda = 3.66$. This TSR range covers the deep and light dynamic stall categories of a VAWT. The case of $L20H20\alpha15$ denotes $L = 0.2c$, $H = 0.2c$, and $\alpha = 15°$.

3.1. $\lambda = 1.75$

Figure 5 illustrates the impact of the BF on the torque of a single blade over one rotation period at $\lambda = 1.75$. The blades experience significant dynamic stall due to a dramatic change in the local angle of attack (AoA). Evident changes can be observed within the range of $\theta = 0°$ to $100°$ in the upwind position and $\theta = 270°$ to $360°$ in the downwind position. The average torque of the blade with the BF increases by 35.8 %, with the average torque increasing by 5.1 % in the upwind position and 87.8 % in the downwind position.

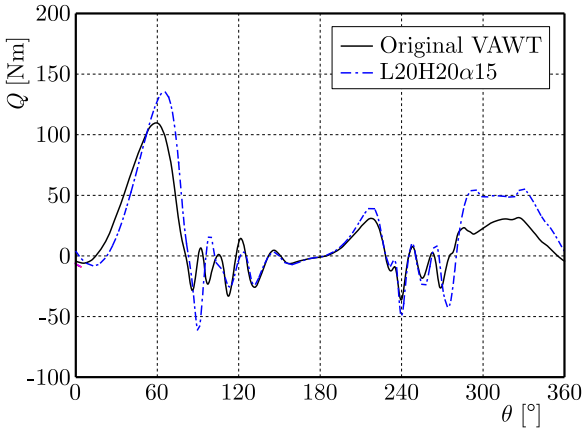

Fig. 5. Torque of a single blade vs. azimuthal angle for $\lambda = 1.75$.

In the upwind position, the torque of the blade with the BF slightly decreases within the range of $\theta = 8°$ to $52°$, where the theoretical local AoA varies from 2.9° to 19.4°. During this

phase, the blade encounters a light dynamic stall, characterized by mostly attached flow or mild flow separation. The presence of the BF restricts the acceleration of flow at the blade's leading edge, resulting in a reduction in lift. With the increase in the azimuthal angle, the local AoA keeps increasing. The torque of the original blade stalls at $\theta = 60°$. As indicated in Fig. 6a, the separation point shifts toward the middle of the blade surface, forming a pair of counter-rotating vortices at the trailing edge. The BF aids in delaying the stall phenomenon. Notably, the torque of the blade with the BF exceeds that of the original blade at $\theta = 52°$ and begins to decrease at $\theta = 66°$. Figure 6b shows that the BF hinders the development of the anticlockwise rotating vortex, resulting in a significantly smaller vortex than that of the original blade, ultimately accelerating the flow at the leading edge. Consequently, the region of negative pressure expands, and its magnitude increases, accounting for the increase in torque from $\theta = 52°$ to 86°. However, after this range, the blade enters a state of deep stall, rendering the BF ineffective.

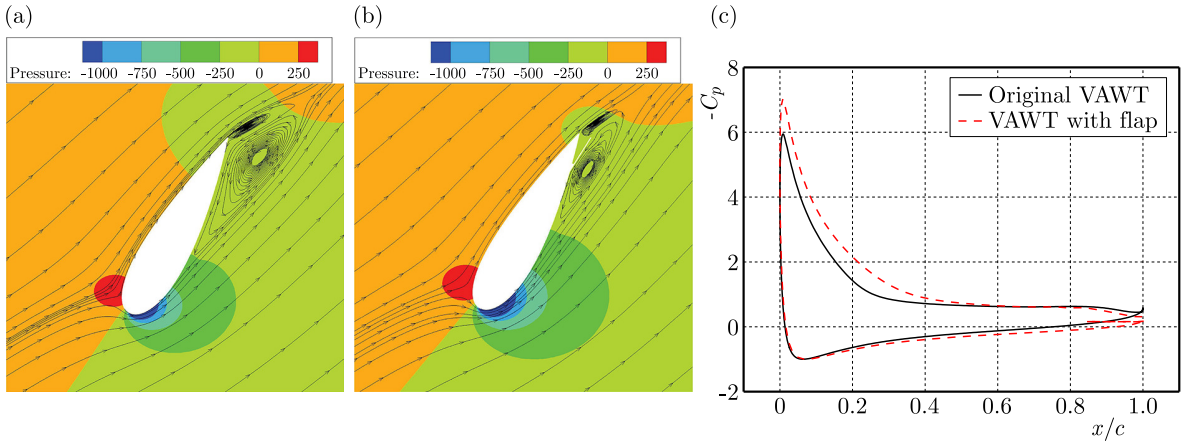

Fig. 6. Flow fields and pressure distributions around the blades at $\theta = 66°$: (a) original blade; (b) blade with the BF; (c) pressure distribution.

In the downwind position, the BF significantly enhances the torque during the rotational period. A notable increase in torque is observed within the range of $\theta = 282°$ to 360°, where the local AoA gradually decreases. As depicted in Fig. 7, the flow reattaches to the blade surface as the azimuthal angle progresses. The BF functions similarly to a GF, improving blade torque through two key mechanisms. First, the BF significantly reshapes the flow field at the trailing edge of the blade. The vortex core at the trailing edge of the blade with the BF is displaced downstream, enhancing the camber effect. This alteration hastens the flow velocity at the leading edge, thereby increasing the negative pressure on the outer surface of the blade with the BF. Second, flow over the inner surface of the blade with the BF experiences blockage and deceleration, due to the changed core position of the trailing-edge vortex and the raised height of the BF. Consequently, the pressure on the inner surface of the blade with the BF goes up. The combined effects of these two mechanisms lead to an increased pressure differential between the inner and outer surfaces of the blade, ultimately enhancing the lift generated by the blade with the BF.

3.2. $\lambda = 3.31$

Figure 8 illustrates the impact of the BF on the torque of a single blade during a rotation period at $\lambda = 3.31$. The blades exhibit light dynamic stall. The torque fluctuations depicted in Fig. 8 are noticeably reduced compared with those observed at $\lambda = 1.75$. The blade with the BF demonstrated an average torque increment of 7.7 % as opposed to the original blade. However, the average torque of the blade with the BF drops by 16.0 % when in the upwind position. This decline in torque is primarily evident within the range of $\theta = 15°$ to 103°. The theoretical local

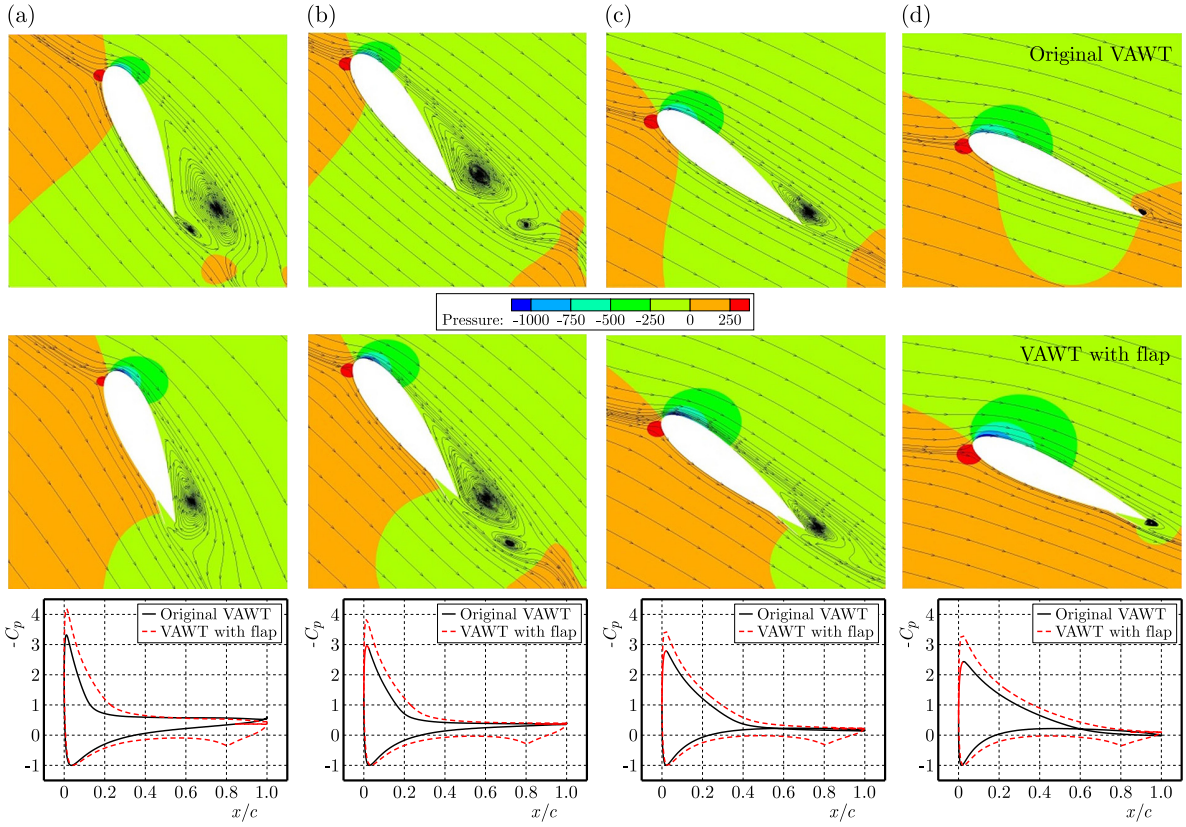

Fig. 7. Flow fields and pressure distributions around the blades at:
(a) $\theta = 288°$, (b) $\theta = 300°$, (c) $\theta = 315°$, (d) $\theta = 330°$.

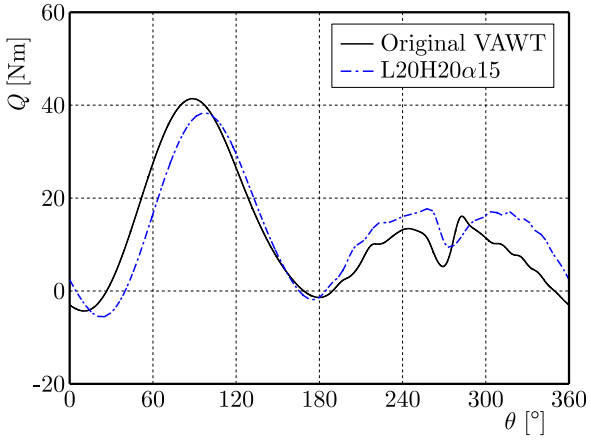

Fig. 8. Torque of a single blade vs. azimuthal angle for $\lambda = 3.31$.

AoA falls between 3.5° and 17.5°. The reasoning behind this phenomenon aligns with that at $\lambda = 1.75$.

The increase in torque for the blade with the BF is attributed entirely to the downwind position, where the average torque increases by 60.4 % as compared to the original blade. Flow attachment is consistently maintained in the downwind positions. As demonstrated in Fig. 9, the BF redirects the flow near the trailing edge, altering the Kutta condition and enhancing the flow circulation around the blade. Furthermore, a pair of counter-rotating vortices forms on the inner surface of the trailing edge of the blade with the BF. The shedding of these vortices increases the suction force on the outer surface of the blade while increasing the pressure on the inner surface by decelerating the flow.

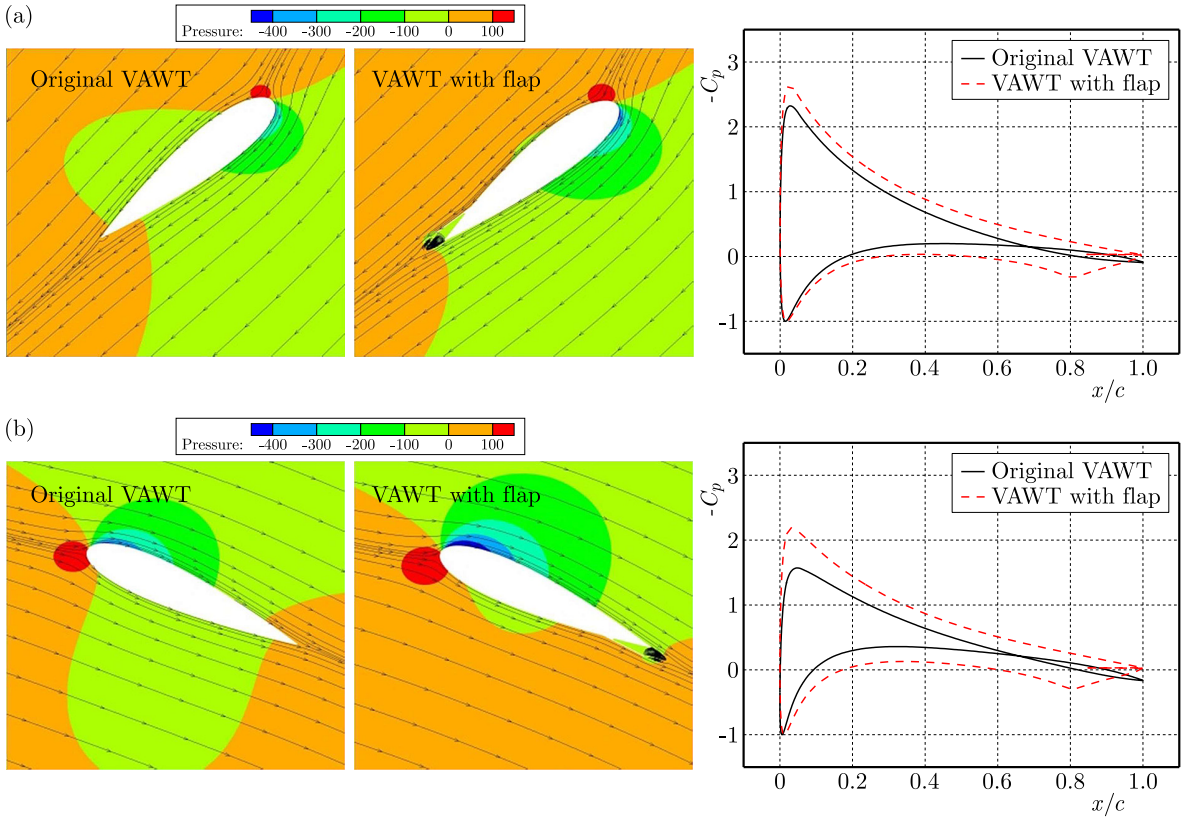

Fig. 9. Flow fields and pressure distributions around the blades at: (a) $\theta = 225°$; (b) $\theta = 330°$.

4. Effects of the BF geometry parameters

The primary geometrical parameters of the BF, as illustrated in Fig. 1, significantly influence the aerodynamic performance of the VAWT. This section discusses the effects of these parameters on the power coefficient of the VAWT and the torque of a single blade.

4.1. Pop-up angle of the BF

A contrast of the power coefficients of the original VAWT and the VAWT with the BF at varying pop-up angles (α) is depicted in Fig. 10. The parameters L and H are both fixed at $0.2c$. The power coefficients are significantly enhanced by the adding of the BF when the TSR $\lambda \leq 3.06$. A relatively large pop-up angle is beneficial for improving the power coefficient

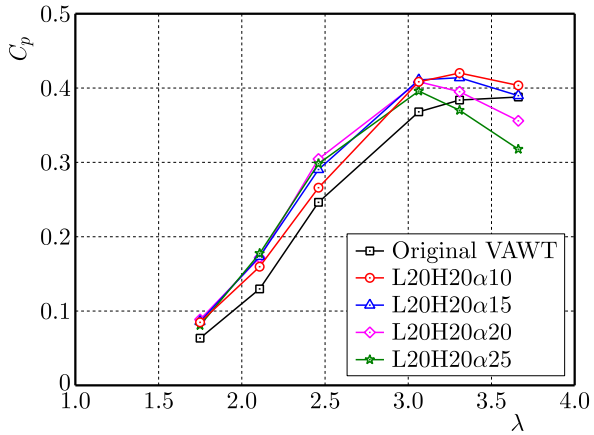

Fig. 10. Power coefficient as a function of TSR at different pop-up angles.

at lower TSRs ($\lambda < 3.06$), with a maximum relative increment of 39.6 % observed at $\lambda = 1.75$ for $\alpha = 20°$.

Figure 11a presents the torques of a single blade as a function of the azimuthal angle for different pop-up angles at $\lambda = 1.75$. Notably, a greater pop-up angle hinders the acceleration of the flow along the inner surface of the blade within the range of $\theta = 8°$ to 60°. The case with $\alpha = 25°$ results in the most significant decrease in torque. However, in the range of $\theta = 60°$ to 86°, a larger pop-up angle delivers superior performance, as the BF efficiently suppresses the trailing-edge vortex. In the downwind position, the relative increments in the average torque of the blade with the BF as compared to the original blade are 76.5 %, 87.8 %, 107.1 %, and 103.7 %, respectively.

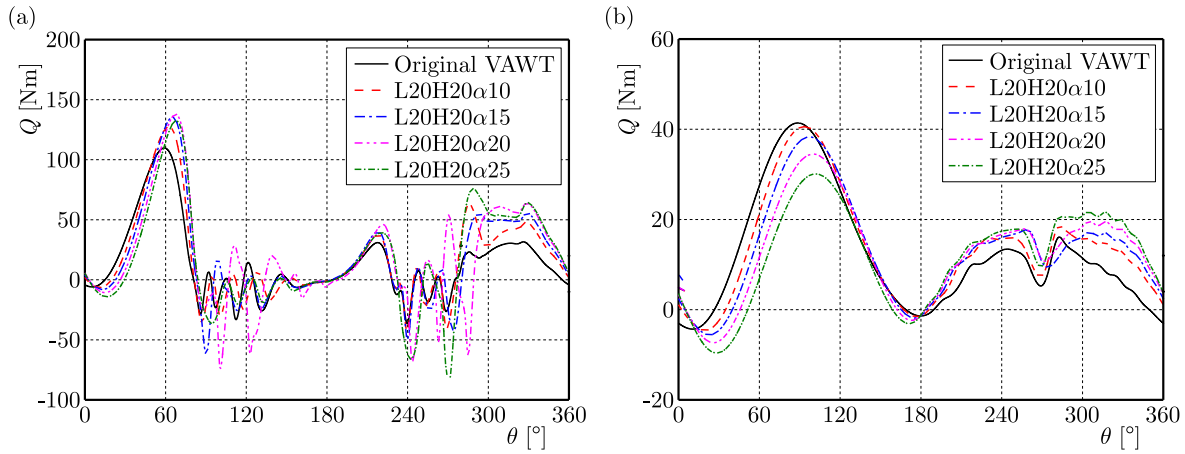

Fig. 11. Torque of a single blade vs. azimuthal angle for different pop-up angles: (a) $\lambda = 1.75$; (b) $\lambda = 3.31$.

As depicted in Fig. 12, the variations in pressure distribution among these cases at $\lambda = 1.75$ are primarily concentrated on the leading edge of the blade's outer surface and the trailing edge of the blade's inner surface. A larger pop-up angle promotes the flow speedup near the leading edge on the outer surface and slows down the flow near the trailing edge on the inner surface. Thus, the negative pressure at the leading edge and the positive pressure at the trailing edge rise as the pop-up angle enlarges. However, a relatively large pop-up angle also induces additional fluctuations in torque, which can negatively affect the fatigue life of the blade.

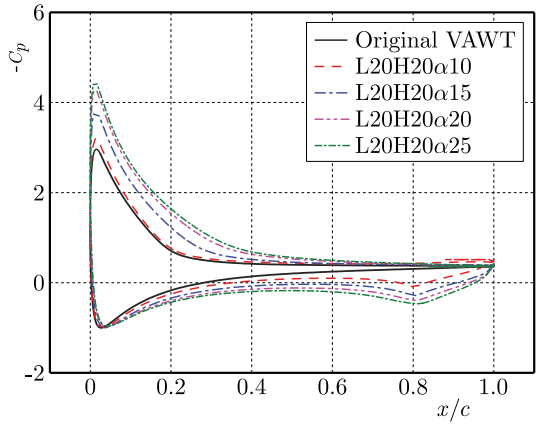

Fig. 12. Pressure distributions around the blades at $\theta = 300°$ for different pop-up angles.

In contrast, a relatively small pop-up angle is beneficial for enhancing the power coefficient at a TSR of $\lambda \geq 3.06$. Specifically, the case with an angle of $\alpha = 10°$ has the highest power coefficient within the range of $\lambda = 3.06$ to 3.66. Compared with those of the original VAWT, the relative increases in the power coefficient are 11.0 %, 9.5 %, and 4.0 % for λ values of 3.06,

3.31, and 3.66, respectively. Notably, the relative increment in the power coefficient diminishes as the pop-up angle increases. However, for a pop-up angle of $\alpha = 25°$, the power coefficient of the VAWT with a BF becomes lower than that of the original VAWT at $\lambda = 3.31$, with a further reduction observed at $\lambda = 3.66$. Compared with the original VAWT, the relative decrements in the power coefficient at $\lambda = 3.66$ are 8.3 % and 18.1 % for the cases with $\alpha = 20°$ and $\alpha = 25°$, respectively.

Figure 11b shows the torques of a single blade versus the azimuthal angle for various pop-up angles at $\lambda = 3.31$. The torque of the VAWT with the BF clearly decreases as the pop-up angle increases in the upwind position. Conversely, the opposite trend is observed in the downwind direction. However, the improvement in torque during the downwind position does not mitigate the reduction experienced in the upwind position for the case with $\alpha = 25°$. Consequently, a relatively small pop-up angle is recommended for enhancing the power coefficient of a VAWT with a BF at a wider TSR range.

For an adjustable BF, the pop-up angle can be tailored on the basis of either the TSR or the azimuthal angle. By tuning the pop-up angle in accordance with the TSR, the power coefficient of the VAWT with an adjustable BF can be determined by selecting the maximum value at each TSR, as shown in Fig. 10. If the pop-up angle is modified on the basis of the azimuthal angle, the BF remains retracted at the azimuthal angles where the flap negatively affects the torque. Two optimal control strategies are formulated on the basis of the findings in Fig. 11 for $\lambda = 1.75$ and 3.31. As depicted in Fig. 13, the BF remains retracted within the range of $\theta = 0°$ to 53° while extending to a pop-up angle of 15° in the range of $\theta = 53°$ to 360° at $\lambda = 1.75$. Compared with the original VAWT, the power coefficient of the VAWT with the BF is enhanced by 45.2 %, which is 5.6 % greater than that of the $L20H20\alpha20$ configuration. For $\lambda = 3.31$, the BF remains retracted in the upwind position while extending to a pop-up angle of 25° in the downwind position. Compared with that of the original VAWT, the power coefficient of the VAWT with the BF increases by 28.9 %, which is 19.4 % greater than that of the $L20H20\alpha10$ configuration. Thus, the adjustable BF demonstrates superior performance in improving the power coefficient relative to a fixed BF. However, the above inference was based on the results of the BF with a fixed pop-up angle. An adjustable BF may influence the development of the vortices and the pressure distribution, which affects the torque history and the power coefficient. Further studies are warranted to reveal the effect of an adjustable BF on the flow separation of the VAWT blade.

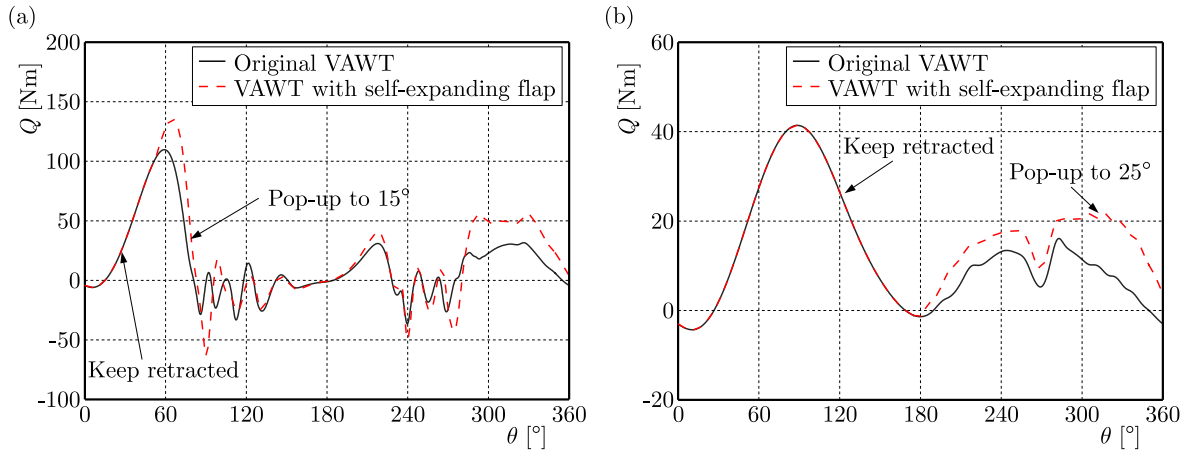

Fig. 13. Torque of a single blade for the VAWT with an adjustable BF: (a) $\lambda = 1.75$; (b) $\lambda = 3.31$.

4.2. Length of the BF

Figure 14 shows the relationship between the length of the BF and the power coefficient, with parameters H and α holding constant at $H = 0.2c$ and $\alpha = 20°$. The variations in the power coefficient caused by changes in BF length are analogous to those induced by alterations in the

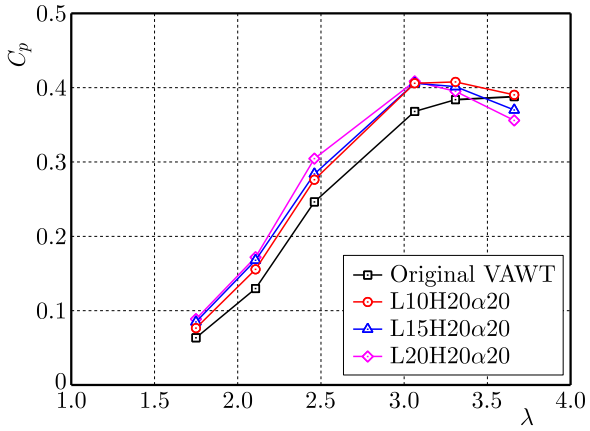

Fig. 14. Power coefficient as a function of TSR at different lengths.

pop-up angle. For $\lambda < 3.06$, the VAWT with a longer BF results in a higher power coefficient. For $\lambda \geq 3.06$, the BF with $L = 0.1c$ outperforms the other two configurations.

4.3. Hinge position of the BF

The hinge position of the BF significantly influences the power coefficient of the VAWT and the torque of a single blade. As shown in Fig. 15, parameters L and α remain constant at $L = 0.2c$ and $\alpha = 20°$. The introduction of the BF leads to substantial increases in power coefficients for $\lambda \leq 3.06$, with a maximum relative increment of 41.3 % observed at $\lambda = 2.11$ with $H = 0.3c$. No discernible trends emerge regarding the effect of the hinge position on the power coefficient of the VAWT with the BF for $\lambda < 3.06$. However, the flap with $H = 0.25c$ demonstrates a smaller improvement in the power coefficient than the other two configurations do. At $\lambda \geq 3.06$, shifting the hinge point of the BF to the blade's leading edge culminates in a diminished power coefficient. The most significant relative reduction is 23.3 % at $\lambda = 3.66$ for $H = 0.3c$.

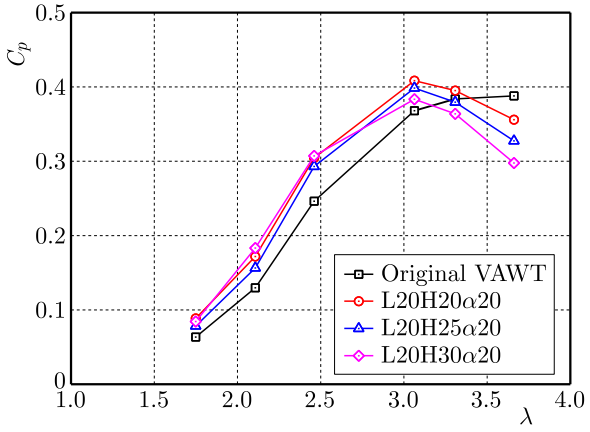

Fig. 15. Power coefficient as a function of TSR at different hinge positions.

5. Conclusions

Numerical simulations are conducted to elucidate the mechanisms by which the BF alleviates flow separation in the VAWT blade and its resultant effects on the power coefficient of the VAWT. The conclusions are as follows:

1) Different mechanisms are exhibited by the BF on blade flow control for the upwind and downwind positions. A BF attached to the inner surface of the blade impedes the development of the trailing-edge vortex at large AoAs when the blade is operating in the upwind

position. Compared with the original blade, the dimensions of the trailing-edge vortex are substantially diminished, which enhances the flow speedup at the leading edge of the blade with a BF. However, the BF impedes flow acceleration at the leading edge at low AoA and detracts from the blade torque at high TSRs. In contrast, when the blade operates in the downwind position, the BF functions similarly to a GF and contributes most to the torque increment, making the BF advantageous for improving torque in the VAWT across a broad range of TSRs.

2) The BF positioned on the inner surface of the blade significantly enhances the power coefficient when $\lambda \leq 3.06$. As the TSR increases further, its effect on power coefficient improvement diminishes due to the decrease in the local AoA. Additionally, the influence of the BF on the power coefficient is contingent upon the three geometric parameters. According to the parameter study, it is recommended to install the BF near the trailing edge of the blade, a hinge position $H < 0.2c$ and the pop-up angle should be smaller than 15°. A multi-objective optimization is needed for improving the power coefficient for a wider range of TSR.

3) Compared with a fixed BF, an adjustable BF, tailored on the basis of the azimuthal angle, has greater potential for improving the power coefficient. Compared with those of the original VAWT, the power coefficients of the VAWT with an adjustable BF are elevated by 45.2 % at $\lambda = 1.75$ and by 28.9 % at $\lambda = 3.31$, significantly outpacing the performance of the VAWT with a fixed BF. Further in-depth studies are necessary to develop a control strategy for the pop-up action of the flap.

Acknowledgments

The Natural Science Foundation of Jiangxi Province (grant no. 20224BAB214061 and 20242BAB20217), the National Natural Science Foundation of China (grant no. 52166002).

References

1. Abbasi, S., & Daraee, M.A. (2024). Ameliorating a vertical axis wind turbine performance utilizing a time-varying force plasma actuator. *Scientific Reports*, *14*, Article 18425. https://doi.org/10.1038/s41598-024-69455-8

2. Abdolahifar, A., & Zanj, A. (2025). A review of available solutions for enhancing aerodynamic performance in Darrieus vertical-axis wind turbines: A comparative discussion. *Energy Conversion and Management*, *327*, Article 119575. https://doi.org/10.1016/j.enconman.2025.119575

3. Attie, C., ElCheikh, A., Nader, J., & Elkhoury, M. (2022). Performance enhancement of a vertical axis wind turbine using a slotted deflective flap at the trailing edge. *Energy Conversion and Management*, *273*, Article 116388. https://doi.org/10.1016/j.enconman.2022.116388

4. Chang, H., Li, D., Zhang, R., Wang, H., He, Y., Zuo, Z., & Liu, S. (2024). Effect of discontinuous biomimetic leading-edge protuberances on the performance of vertical axis wind turbines. *Applied Energy*, *364*, Article 123117. https://doi.org/10.1016/j.apenergy.2024.123117

5. Chen, L., Xu, J., Dai, R. (2020). Numerical prediction of switching gurney flap effects on straight bladed VAWT power performance. *Journal of Mechanical Science and Technology*, *34*(12), 4933–4940. https://doi.org/10.1007/s12206-020-2106-z

6. Daróczy, L., Janiga, G., Petrasch, K., Webner, M., & Thévenin, D. (2015). Comparative analysis of turbulence models for the aerodynamic simulation of H-Darrieus rotors. *Energy*, *90*(Part 1), 680–690. https://doi.org/10.1016/j.energy.2015.07.102

7. Ertem, S., Ferreira, C.S., Gaunaa, M., & Madsen, H.A. (2016). Aerodynamic optimization of vertical axis wind turbine with trailing edge flaps. *34th Wind Energy Symposium, AIAA 2016-1735*. https://doi.org/10.2514/6.2016-1735

8. Han, Z., Chen, H., Chen, Y., Su, J., Zhou, D., Zhu, H., Xia, T., & Tu, J. (2023). Aerodynamic performance optimization of vertical axis wind turbine with straight blades based on synergic control of pitch and flap. *Sustainable Energy Technologies and Assessments*, *57*, Article 103250. https://doi.org/10.1016/j.seta.2023.103250

9. Kjellin, J., Bülow, F., Eriksson, S., Deglaire, P., Leijon, M., & Bernhoff, H. (2011). Power coefficient measurement on a 12 kW straight bladed vertical axis wind turbine. *Renewable Energy*, *36*(11), 3050–3053. https://doi.org/10.1016/j.renene.2011.03.031

10. Liu, Q., Miao, W., Ye, Q., & Li, C. (2022). Performance assessment of an innovative Gurney flap for straight-bladed vertical axis wind turbine. *Renewable Energy*, *185*, 1124–1138. https://doi.org/10.1016/j.renene.2021.12.098

11. Ma, Q., Wang, J., Zhang, Y., Ma, L., Lin, J., & Liu, X. (2022). Research on optimized design of structural parameters and aerodynamics of wind turbine airfoil with bionic flap (in Chinese). *Journal of Xi'an Jiaotong University*, *56*(11), 31–40.

12. McKenna, R., Lilliestam, J., Heinrichs, H.U., Weinand, J., Schmidt, J., Staffell, I., Hahmann, A.N., Burgherr, P., Burdack, A., Bucha, M., Chen, R., Klingler, M., Lehmann, P., Lowitzsch, J., Novo, R., Price, J., Sacchi, R., Scherhaufer, P., Schöll, E.M., Visconti, P., Camargo, L.R. (2025). System impacts of wind energy developments: Key research challenges and opportunities. *Joule*, *9*(1), Article 101799. https://doi.org/10.1016/j.joule.2024.11.016

13. Patankar, S.V., & Spalding, D.B. (1972). A calculation procedure for heat, mass and momentum transfer in three-dimensional parabolic flows. *International Journal of Heat and Mass Transfer*, *15*(10), 1787–1806. https://doi.org/10.1016/0017-9310(72)90054-3

14. Syawitri, T.P., Yao, Y., Yao, J., & Chandra, B. (2024). Drag reduction of lift-type Vertical axis wind turbine with slit modified Gurney flap. *Journal of Wind Engineering and Industrial Aerodynamics*, *253*, Article 105853. https://doi.org/10.1016/j.jweia.2024.105853

15. Rehman, S., Alhems, L.M., Alam, M.M., Wang, L., & Toor, Z. (2023). A review of energy extraction from wind and ocean: Technologies, merits, efficiencies, and cost. *Ocean Engineering*, *267*, Article 113192. https://doi.org/10.1016/j.oceaneng.2022.113192

16. Rezaeiha, A., Montazeri, H., & Blocken, B. (2018). Towards accurate CFD simulations of vertical axis wind turbines at different tip speed ratios and solidities: Guidelines for azimuthal increment, domain size and convergence. *Energy Conversion and Management*, *156*, 301–316. https://doi.org/10.1016/j.enconman.2017.11.026

17. Rezaeiha, A., Montazeri, H., & Blocken, B. (2019). Active flow control for power enhancement of vertical axis wind turbines: Leading-edge slot suction. *Energy*, *189*, Article 116131. https://doi.org/10.1016/j.energy.2019.116131

18. Rogowski, K., Hansen, M.O.L., & Maronski, R. (2018). Steady and unsteady analysis of NACA 0018 airfoil in vertical-axis wind turbine. *Journal of Theoretical and Applied Mechanics*, *56*(1), 203–212. https://doi.org/10.15632/jtam-pl.56.1.203

19. Sridhar, S., Joseph, J., & Radhakrishnan, J. (2022). Implementation of tubercles on Vertical Axis Wind Turbines (VAWTs): An aerodynamic perspective. *Sustainable Energy Technologies and Assessments*, *52*(Part B), Article 102109. https://doi.org/10.1016/j.seta.2022.102109

20. Sun, J., & Huang, D. (2023). Impact of trailing edge jet on the performance of a vertical axis wind turbine. *Journal of Mechanical Science and Technology*, *37*(3), 1301–1309. https://doi.org/10.1007/s12206-023-0216-0

21. Tayebi, A., & Torabi, F. (2024). Flow control techniques to improve the aerodynamic performance of Darrieus vertical axis wind turbines: A critical review. *Journal of Wind Engineering and Industrial Aerodynamics*, *252*, Article 105820. https://doi.org/10.1016/j.jweia.2024.105820

22. Ullah, T., Javed, A., Abdullah, A., Ali, M., & Uddin, E. (2020). Computational evaluation of an optimum leading-edge slat deflection angle for dynamic stall control in a novel urban-scale vertical axis wind turbine for low wind speed operation. *Sustainable Energy Technologies and Assessments*, *40*, Article 100748. https://doi.org/10.1016/j.seta.2020.100748

23. Yan, Y., Avital, E.J., Williams, J., & Cui, J. (2019). CFD analysis for the performance of micro-vortex generator on aerofoil and vertical axis turbine. *Journal of Renewable and Sustainable Energy*, *11*(4), Article 043302. https://doi.org/10.1063/1.5110422

24. Yan, Y., Avital, E.J., Williams, J., & Cui, J. (2021). Aerodynamic performance improvements of a vertical axis wind turbine by leading-edge protuberance. *Journal of Wind Engineering and Industrial Aerodynamics*, *211*, Article 104535. https://doi.org/10.1016/j.jweia.2021.104535

25. Zhao, Z., Wang, D., Wang, T., Shen, W., Liu, H., & Chen, M. (2022). A review: Approaches for aerodynamic performance improvement of lift-type vertical axis wind turbine. *Sustainable Energy Technologies and Assessments*, *49*, Article 101789. https://doi.org/10.1016/j.seta.2021.101789

26. Zhong, J., Li, J., Guo, P., & Wang, Y. (2019). Dynamic stall control on a vertical axis wind turbine aerofoil using leading-edge rod. *Energy*, *174*, 246–260. https://doi.org/10.1016/j.energy.2019.02.176

27. Zhu, H., Hao, W., Li, C., Luo, S., Liu, Q., & Gao, C. (2021). Effect of geometric parameters of Gurney flap on performance enhancement of straight-bladed vertical axis wind turbine. *Renewable Energy*, *165*(Part 1), 464–480. https://doi.org/10.1016/j.renene.2020.11.027

Manuscript received December 3, 2024; accepted for publication April 3, 2025; published online October 1, 2025.

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 839–853, Warsaw 2025
https://doi.org/10.15632/jtam-pl/203694

A FAST AND ROBUST NUMERICAL SOLUTION FOR THE POSITIONING DESIGN OF HORIZONTAL DRAINS IN A SATURATED-UNSATURATED SOIL GROUND SYSTEM

Qian HE*, Quan YUAN, Lin LANG

School of Architecture and Civil Engineering, Xihua University, Chengdu, China
*corresponding author, *qianhxhu@163.com*

This paper proposes an implicit difference solution to design the position of the horizontal drain in a saturated-unsaturated soil ground system (SUSGS). A complete system of classic diffusion equations is used to describe the transient flow of air and water phases in the soil system. Consolidation equations and boundary conditions are discretized by using the Crack–Nicolson (C-N) and virtual grid methods, respectively. The numerical scheme has been verified to be unconditionally stable based on von Neumann's theorem. The comparison with existing analytical predictions confirms that the proposed solution is effective and accurate. According to the verified numerical solution, the optimum position of horizontal drains is designed to elevate the consolidation rate of the saturated-unsaturated soil ground system.

Keywords: consolidation; implicit difference solution; horizontal drain; saturated-unsaturated system; virtual grid method.

1. Introduction

In nature, the soil above and below the groundwater level exhibits unsaturated and saturated states, which forms a saturated-unsaturated soil ground system (SUSGS) (Li *et al.*, 2021). Due to the infiltration of rainfall, transpiration of plants, and extraction of groundwater, the soil system will change with the phreatic line. The decline or rise of groundwater may lead to building crack and non-uniform settlement of grounds. Therefore, some engineering requirements have been proposed to accelerate the discharge of the water phase. For instance, the horizontal drain is widely used in the construction of large airports (Mesri & Funk, 2015), highways (Zhou *et al.*, 2023), and railways (Gu *et al.*, 2020) to accelerate the drainage consolidation and improve the bearing capacity of the ground. The horizontal drain is placed between soil layers to shorten the seepage path of fluid and increase the consolidation rate of soil ground (Gibson & Shefford, 1968; Zhou *et al.*, 2024). Thus, it is meaningful to preliminarily assess the consolidation behavior of the SUSGS with horizontal drains.

Using the saturated soil model alone in geological engineering to solve drainage consolidation problems is convenient. Classical consolidation theory established by Terzaghi (1943) was applied to predict the consolidation behavior of the saturated soil. Some extended models are intended for the actual engineering problems, such as the ground with the horizontal drain (Feng *et al.*, 2019), layered soil (Feng *et al.*, 2020), and non-Darcy flow (Wu *et al.*, 2023). Based on Terzaghi's theory and the engineering practice, a series of engineering technologies have been proposed to enhance the consolidation rate of soil ground. The horizontal drain is embedded at equal intervals in the soil layers to accelerate the consolidation rate in some geotechnical engineering (Lee *et al.*, 1987; Mesri & Funk, 2015). Meng *et al.* (2019), Li *et al.* (2020), and Feng *et al.* (2020) investigated the consolidation behavior of single-layer, double-layer, and four-layer saturated ground with the

horizontal drains. They provided constructive design schemes for the horizontal drain under different soil parameters.

For the consolidation of unsaturated soil, two typical diffusion equations for the air and water phases were obtained by Fredlund and Hasan (1979), which was used to estimate the dissipation of excess pore pressures and settlement of unsaturated soil. This original model and its extended form have been applied in various geotechnical engineering issues (Wang *et al.*, 2019; Yuan *et al.*, 2023; He *et al.*, 2025). The consolidation of unsaturated soil with horizontal drains needs more consideration. Zhou *et al.* (2023; 2024) proposed a semi-analytical solution to predict the consolidation behavior of double-layer unsaturated soil ground and a double-layer saturated-unsaturated soil ground system with the horizontal drain. However, in Zhou's literature, the horizontal drain is located at the interface between the upper and lower soil layers, and the position of the horizontal drain is not improved to be the best for discharging water. It is important to notice that the explicit difference method is employed to verify the correctness of analytical or semi-analytical solutions for the positioning design of horizontal drains in the SUSGS (Qin *et al.*, 2008; Wang *et al.*, 2017; 2019). With conditional stability and slow convergence, this difference scheme will cause high computational costs and poor reliability in solving consolidation problems.

The study reported in this paper attempts to develop an implicit difference solution with high computational efficiency and accuracy to design the position of horizontal drains in the SUSGS. The diffusion equations and boundary conditions are discretized using the Crack–Nicolson (C-N) and virtual grid methods. Then, the distribution of excess pore-air and pore-water pressures is obtained through matrix operation. Comparisons with the existing solutions are performed to verify the reliability and effectiveness of the numerical solution. Based on the proposed solution, the influence of horizontal drains on the SUSGS has been investigated, and the optimizing design of depths of horizontal drains is provided.

2. Mathematical model

2.1. Model description and basic assumptions

Combining the consolidation theories of the saturated and unsaturated soils proposed by Terzaghi (1943) and Fredlund *et al.* (2012), a simple mathematical description of the consolidation behavior of the SUSGS with horizontal drains is shown in Fig. 1. The SUSGS is divided into four zones by the embedment of two horizontal drains and an interface. Zone I comprises unsat-

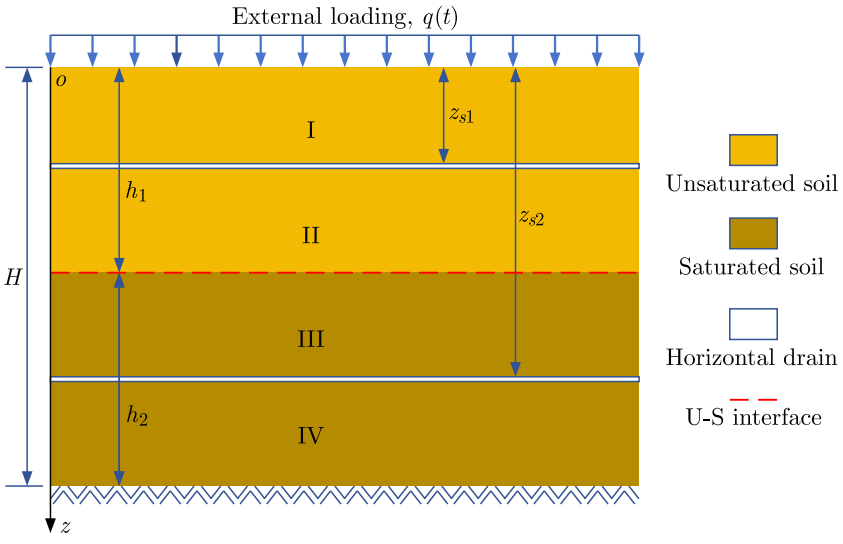

Fig. 1. Schematic diagram of SUSGS with horizontal drains.

urated soil with a two-way drainage system; the top and lower surfaces are permeable to air and water. In zone II, the upper surface is permeable to air and water phases, and the lower surface is impermeable to the air phase. Zone III is composed of saturated soil with a two-way drainage system. Zone IV has a one-way drainage system; the upper surface is permeable to water, and the bottom surface is impermeable. In the ground system, the thicknesses of the unsaturated and saturated soils are h_1 and h_2, respectively, $h_1+h_2=\mathrm{H}$; z_{s1} and z_{s2} are the imbedded depths of the first and second horizontal drains, respectively.

The main assumptions of the consolidation model for the SUSGS are as follows (Li *et al.*, 2021): (1) the unsaturated soil layer and saturated soil layer are both homogeneous; (2) the soil particles, water phase and horizontal drain are incompressible; (3) the flow of air and water phases along z-direction are independent, linear, and continuous steady-state; (4) in the consolidation process, the consolidation parameters keep constant; (5) the thickness of the capillary water zone is very thin, and its impact on the hydraulic response of the SUSGS is not significant.

It is worth noting that the above assumptions are not exact for all situations. For example, the soil is not homogeneous; it is composed of particles of different sizes and air and water. The influence of the capillary water zone on soils may be very significant. If we consider all these real situations, analyzing the consolidation behavior of such a complex system is far beyond our capability. In order to simplify the mathematical derivation process and preliminarily predict the consolidation behavior of the SUSGS, these listed assumptions are essential for developing the analytical and numerical solutions for the consolidation equations (Moradi *et al.*, 2019; Yuan *et al.*, 2023).

2.2. Consolidation equations

The volume changes associated with the water and air phases during the consolidation process can be calculated using Darcy's and Fick's laws. Then, the volume changes of the soil unit are equal to the sum of the volume changes of the air and water, and the governing equations are obtained (Li *et al.*, 2022):

$$\mathbf{C}\frac{\partial \mathbf{u}}{\partial t}+\mathbf{Z}\frac{\partial^2 \mathbf{u}}{\partial z^2}=\frac{\partial \mathbf{q}}{\partial t}, \qquad (0<z<h_1), \tag{2.1}$$

$$\frac{\partial u_v}{\partial t}+c_{vz}\frac{\partial^2 u_v}{\partial z^2}=\frac{\partial q}{\partial t}, \qquad (h_1<z<H), \tag{2.2}$$

where

$$\mathbf{u}=\begin{bmatrix} u_a \\ u_w \end{bmatrix}, \qquad \mathbf{C}=\begin{pmatrix} 1 & c_a \\ c_w & 1 \end{pmatrix}, \qquad \mathbf{Z}=\begin{pmatrix} c_{vz}^a & 0 \\ 0 & c_{vz}^w \end{pmatrix}, \qquad \mathbf{q}=\begin{bmatrix} c_\sigma^a \\ c_\sigma^w \end{bmatrix} q(t),$$

u_a and u_w represent excess pore-air and pore-water pressures of unsaturated soil, respectively, u_w is excess pore-water pressure of saturated soil. In Eq. (2.1), c_a, c_{vz}^a, and c_σ^a are consolidation parameters of the air phase, c_w, c_{vz}^w, and c_σ^w are consolidation parameters of the water phase. In Eq. (2.2), c_{vz} is consolidation parameters of the saturated soil. The definitions and expressions of these consolidation parameters can be found in (Fredlund, 2012; Terzaghi, 1943).

2.3. Model conditions

The drainage conditions of the SUSGS are:
– for zone I

$$\mathbf{u}(0,t)=\mathbf{0}, \tag{2.3}$$

$$\mathbf{u}(z_{s1},t)=\mathbf{0}, \tag{2.4}$$

– for zone II

$$\mathbf{u}(z_{s1},t)=\mathbf{0}, \tag{2.5}$$

$$\frac{\partial u_a(h_1,t)}{\partial z}=0, \tag{2.6}$$

$$k_w\frac{\partial u_w(h_1,t)}{\partial z}=k_v\frac{\partial u_v(h_1,t)}{\partial z},\qquad u_w(h_1,t)=u_v(h_1,t), \tag{2.7}$$

– for zone III

$$k_w\frac{\partial u_w(h_1,t)}{\partial z}=k_v\frac{\partial u_v(h_1,t)}{\partial z},\qquad u_w(h_1,t)=u_v(h_1,t) \tag{2.8}$$

$$u_v(z_{s2},t)=0, \tag{2.9}$$

– for zone IV

$$u_v(z_{s2},t)=0, \tag{2.10}$$

$$\frac{\partial u_v(H,t)}{\partial z}=0. \tag{2.11}$$

The initial conditions are expressed as

$$\mathbf{u}(z,0)=f(z)\left[u_a^0\ \ u_w^0\right]^{\mathrm{T}},\qquad (0<z\le h_1), \tag{2.12}$$

$$u_v(z,0)=f(z)p_0,\qquad (h_1<z<H), \tag{2.13}$$

where u_a^0, u_w^0, and p_0 are initial excess pore pressures at $t=0$; $f(z)$ is a distribution function about z.

3. Solution derivation

3.1. Crack–Nicolson (C-N) solution for consolidation equations

The C-N method has been widely used to solve a single Navier–Stokes equation (Feng *et al.*, 2020). For a system of partial differential equations containing three variables (u_a, u_w, and u_v) and a series of definite solution conditions (Eqs. (2.3)–(2.13)), some technical improvements are needed when using the C-N method to solve it. The difference mesh for the SUSGS is shown in Fig. 2. The time and spatial domains are divided into N and K equidistant nodes, respectively. Time step and space step are τ and h. The node coordinates are (z_k,t_n), and $z_k=kh_z$, $k=0,1,2,...,K$; $t_n=n\tau$, $n=0,1,2,...,N$. z_{k1} and z_{k2} represent the positions of the first and second horizontal drains, respectively, where $z_0<z_{k1}$, $z_{k2}<z_K$. z_M denotes the U-S interface.

Equations (2.1) and (2.2) can be discretized as

$$\mathbf{C}\frac{\mathbf{u}_k^{n+1}-\mathbf{u}_k^n}{\tau}+\mathbf{Z}\frac{\delta_z^2\mathbf{u}_k^{n+1}+\delta_z^2\mathbf{u}_k^n}{2h_z^2}=\frac{\mathbf{q}^{n+1}-\mathbf{q}^n}{\tau},\qquad (1\le j\le M-1), \tag{3.1}$$

$$\frac{u_{vk}^{n+1}-u_{vk}^n}{\tau}+c_{vz}\frac{\delta_z^2u_{vk}^{n+1}+\delta_z^2u_{vk}^n}{2h_z^2}=\frac{q^{n+1}-q^n}{\tau},\qquad (M+1\le j\le K-1), \tag{3.2}$$

where $\delta_z^2\mathbf{u}_j^n=\mathbf{u}_{k-1}^n-2\mathbf{u}_k^n+\mathbf{u}_{k+1}^n$, $\delta_z^2u_{vk}^n=u_{v(k-1)}^n-2u_{vk}^n+u_{v(k+1)}^n$.

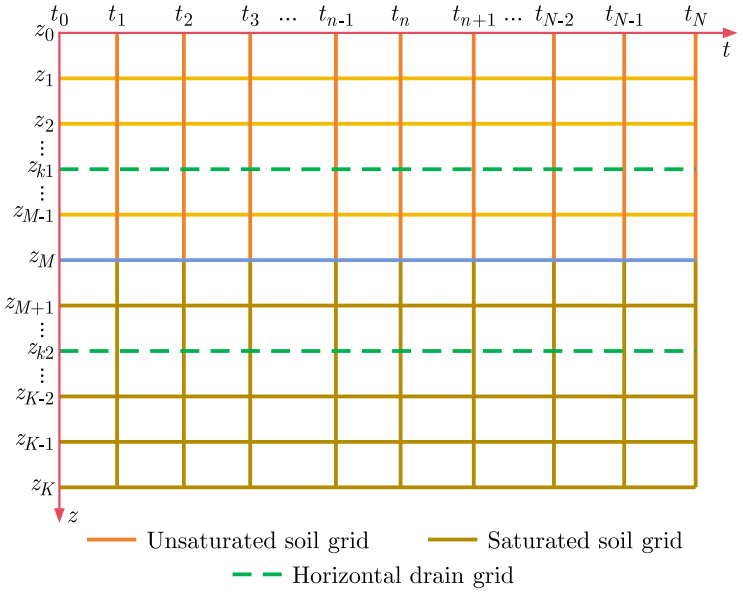

Fig. 2. C-N scheme for difference grid of the saturated-unsaturated soil system ground.

Equations (3.1) and (3.2) can be rewritten as

$$\frac{1}{2}\lambda_z \mathbf{Z}\mathbf{u}_{k-1}^{n+1} + (\mathbf{C}-\lambda_z \mathbf{Z})\,\mathbf{u}_k^{n+1} + \frac{1}{2}\lambda_z \mathbf{Z}\mathbf{u}_{k+1}^{n+1} = -\frac{1}{2}\lambda_z \mathbf{Z}\mathbf{u}_{k-1}^{n} + (\mathbf{C}+\lambda_z \mathbf{Z})\,\mathbf{u}_k^{n} - \frac{1}{2}\lambda_z \mathbf{X}\mathbf{u}_{k-1}^{n} + \mathbf{q}^{n+1} - \mathbf{q}^{n}, \tag{3.3}$$

$$\frac{1}{2}\lambda_z c_{vz} u_{v(k-1)}^{n+1} + (1-\lambda_z c_{vz})\,u_{vk}^{n+1} + \frac{1}{2}\lambda_z c_{vz} u_{v(k+1)}^{n+1} = -\frac{1}{2}\lambda_z c_{vz} u_{v(k-1)}^{n} + (1+\lambda_z c_{vz})\,u_{vk}^{n} - \frac{1}{2}\lambda_z c_{vz} u_{v(k+1)}^{n} + q^{n+1} - q^{n}, \tag{3.4}$$

where $\lambda_z = \tau/h^2$.

Equations (3.3) and (3.4) contain three unknown quantities and three known quantities on adjacent layer k. The value of mesh nodes can only be obtained by using the matrix operation. Equations (3.3) and (3.4) are written as

$$\mathbf{A}_u \mathbf{D}_u^{n+1} = \mathbf{B}_u \mathbf{D}_u^{n} + \mathbf{Q}_u^{n}, \qquad (1 \le j \le M-1)\,, \tag{3.5}$$

$$\mathbf{A}_s \mathbf{D}_s^{n+1} = \mathbf{B}_s \mathbf{D}_s^{n} + \mathbf{Q}_s^{n}, \qquad (M+1 \le j \le K-1)\,, \tag{3.6}$$

where $\mathbf{A}_u$ and $\mathbf{B}_u$ are coefficient matrices, $\mathbf{Q}_u^n$ and $\mathbf{Q}_s^n$ are composed of the external loading, $\mathbf{D}_u^n$ and $\mathbf{D}_s^n$ are composed of the pore-air and pore-water pressures. The specific expressions of these parameters are shown in Appendix A.

It is necessary to validate the stability and convergence of the proposed difference method, which can be verified using von Neumann's theorem (Sharifi & Rashidinia, 2016). The derivation of stability is shown in Appendix B.

3.2. Boundary conditions discretization

The permeable boundary conditions, Eqs. (2.3)–(2.5), (2.9), and (2.10), are regarded as known quantities appearing on the right-hand sides of Eqs. (3.3) and (3.4). If the interface ($z = h_1$ and $z = H$) satisfies the impermeable and continuous permeable boundaries, we use the

virtual mesh method (Morton & Mayers, 2005) to obtain the discrete scheme of Eqs. (2.6)–(2.8), and (2.11).

As shown in Figs. 3a and 3b, the virtual layers of the air and water pressures are constructed outside the boundary. Then, Eqs. (2.6) and (2.11) are discretized by the central difference:

$$\frac{u^n_{v(K+1)} - u^n_{v(K-1)}}{2h_z} = 0, \tag{3.7}$$

$$\frac{u^n_{a(M+1)} - u^n_{a(M-1)}}{2h_z} = 0. \tag{3.8}$$

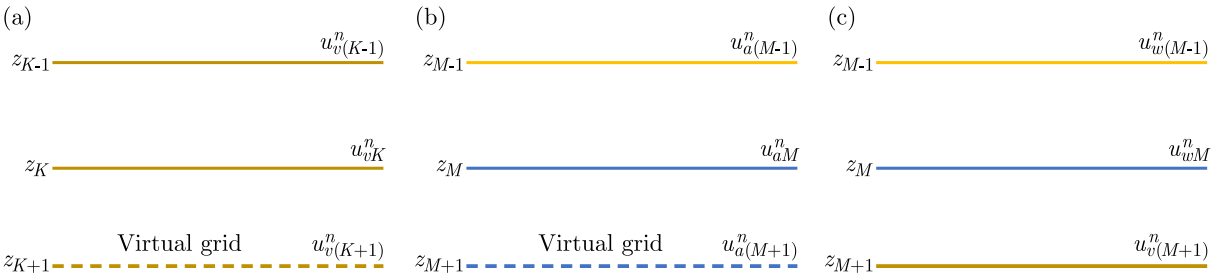

Fig. 3. Difference mesh at the boundary: (a) $z = z_K$; (b) z_M for air phase; (c) z_M for water phase.

Substituting Eqs. (3.7) and (3.8) into Eqs. (3.3) and (3.4) to eliminate the virtual nodes, the difference format at the impermeable boundary is obtained.

The forward and backward difference quotients are used to discretize the left and right sides of Eq. (2.8):

$$k_w \frac{u^n_{wM} - u^n_{w(M-1)}}{h} = k_v \frac{u^n_{vM} - u^n_{v(M+1)}}{h}, \tag{3.9}$$

$$u^n_{wM} = u^n_{vM}. \tag{3.10}$$

As shown in Fig. 3c, the difference scheme of Eq. (3.3) at $z = z_M$ includes the nodes $u^n_{w(M-1)}$, u^n_{wM}, and $u^n_{v(M+1)}$, so it is represented as

$$\begin{aligned}\frac{1}{2}\lambda_z c^w_{vz} u^{n+1}_{w(M-1)} + (1 - \lambda_z c^w_{vz})\, u^{n+1}_{wM} + \frac{1}{2}\lambda_z c_{vz} u^{n+1}_{v(M+1)} + c_w u^{n+1}_{a(M-1)} = -\frac{1}{2}\lambda_z c^w_{vz} u^n_{w(M-1)} \\ + (1 + \lambda_z c^w_{vz})\, u^n_{wM} - \frac{1}{2}\lambda_z c_{vz} u^n_{v(M+1)} + c_w u^n_{a(M-1)} + c^w_\sigma \left(q^{n+1} - q^n\right).\end{aligned} \tag{3.11}$$

In the above equation, $u^{n+1}_{v(M+1)}$ and $u^n_{v(M+1)}$ are the virtual nodes. The difference solution of Eq. (3.3) at $z = z_M$ is obtained by substituting Eq. (3.9) into Eq. (3.11) to eliminate these virtual nodes.

Similarly, the difference solution of Eq. (3.4) at $z = z_M$ contains the nodes $u^n_{w(M-1)}$, u^n_{vM}, and $u^n_{v(M+1)}$ in the iteration process. The difference scheme is expressed as follows:

$$\begin{aligned}\frac{1}{2}\lambda_z c^w_{vz} u^{n+1}_{w(M-1)} + (1 - \lambda_z c_{vz})\, u^{n+1}_{vM} + \frac{1}{2}\lambda_z c_{vz} u^{n+1}_{v(M+1)} = -\frac{1}{2}\lambda_z c^w_{vz} u^n_{w(M-1)} \\ + (1 + \lambda_z c_{vz})\, u^n_{vM} - \frac{1}{2}\lambda_z c_{vz} u^n_{v(M+1)} + q^{n+1} - q^n,\end{aligned} \tag{3.12}$$

where $u^{n+1}_{w(M-1)}$ and $u^n_{w(M-1)}$ are the virtual nodes.

Substituting Eq. (3.9) into Eq. (3.12) to eliminate these virtual nodes, the C-N solution for Eq. (3.4) at $z = z_M$ can be obtained. Adding the difference scheme of Eqs. (3.3) and (3.4) at $z = z_M$, the C-N solution at the interface ($z = h_1$) is

$$\begin{aligned}\alpha_{M-1}u^{n+1}_{w(M-1)} + \alpha_M u^{n+1}_{wM} + \alpha_{M+1}u^{n+1}_{v(M+1)} + c_w u^{n+1}_{a(M-1)} = \beta_{M-1}u^{n}_{w(M-1)} \\ + \beta_M u^{n}_{wM} + \beta_{M+1}u^{n}_{v(M+1)} + c_w u^{n}_{a(M-1)} + (c^w_\sigma + 1)\left(q^{n+1} - q^{n-1}\right),\end{aligned} \tag{3.13}$$

where $\alpha_{m-1}, ..., \gamma_{m+1}$ are constant coefficients.

In the above derivation, we give the difference schemes of the consolidation equations and boundary conditions. The final C-N solution can be expressed as

$$\mathbf{AD}^{n+1} = \mathbf{BD}^n + \mathbf{Q}^n, \tag{3.14}$$

where $\mathbf{A}$ and $\mathbf{B}$ are partitioned matrices. Their description can be found in Appendix C.

3.3. Settlement discretization

Based on generalized Hooke's law, the volume change of soil structure can be formulated by the net normal stress and matrix suction. After obtaining the excess pore pressures by using Eq. (3.14), the volumetric strain of the unsaturated soil is expressed as (Fredlund *et al.*, 2012):

$$d\varepsilon^n_{wk} = m^s_1 d\left(q^n - u^n_{ak}\right) + m^s_2 d\left(u^n_{ak} - u^n_{wk}\right), \tag{3.15}$$

where m^s_1 and m^s_2 are the coefficients of soil volume changes with respect to the net normal stress and matrix suction.

The volumetric strain of the saturated soil is expressed as

$$d\varepsilon^n_{vk} = m_v d\left(q^n - u^n_{vk}\right). \tag{3.16}$$

The settlement of the SUSGS is

$$S\left(t_n\right) = h_z\left(\sum_{k=0}^{M}\varepsilon^n_{wk} + \sum_{k=M+1}^{K}\varepsilon^n_{vk}\right). \tag{3.17}$$

The average degree of consolidation is

$$S^* = \frac{S(t_n)}{S\left(t_\infty\right)} \times 100\,\%, \tag{3.18}$$

where $S\left(t_\infty\right)$ is the maximum settlement of the SUSGS.

4. Verification

In order to verify the convergence (effectiveness and reliability) of the C-N solution, a comparison of the numerical results and analytical solution (Zhou *et al.*, 2023) is made. The consolidation parameters are assumed as follows: $c_a = -0.0899$, $c_w = 0.75$, $c^a_{vx} = -5.3476 \times 10^{-6}\,\mathrm{m/s^2}$, $c^w_{vz} = -5.108 \times 10^{-8}\,\mathrm{m/s^2}$, $u^0_a = 20\,\mathrm{kPa}$, $u^0_w = 40\,\mathrm{kPa}$, $c_{vz} = -8.16310^{-7}\,\mathrm{m/s^2}$, $p_0 = 100\,\mathrm{kPa}$, $H = 10\,\mathrm{m}$, $h_1 = 5\,\mathrm{m}$, $q(t) = 100\,\mathrm{kPa}$. Figure 4 shows the distribution of pore pressures in the SUSGS with one horizontal drain ($t = 10^5\,\mathrm{s}$ for air phase, and $t = 10^7\,\mathrm{s}$ for water phase). The proposed numerical solution agrees with the analytical solutions, suggesting that it is reliable and accurate. Further validation for the computational efficiency of the proposed algorithm under different grid ratios is given in Appendix D.

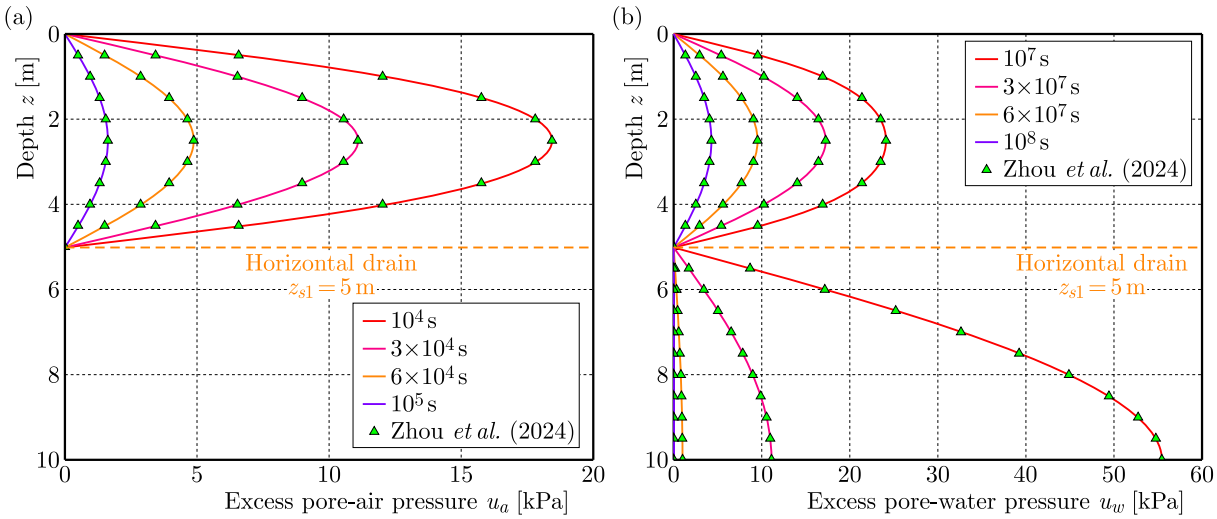

Fig. 4. Distribution of excess pore pressures along the depth direction: (a) u_a and (b) u_w.

5. Numerical examples

5.1. The position design of a single horizontal drain

The purpose of installing the horizontal drain is to accelerate the consolidation rate of the ground. The optimal position for the horizontal drain is to achieve the consolidation degree of 90 % in the shortest time (Meng *et al.*, 2019). Hence, the objective function can be expressed as

$$\min T_v = T_{90}(z_{s1}), \tag{5.1}$$

where T_v is the objective function, T_{90} is the time cost for consolidation degree of 90 %.

Based on Eq. (3.18), the relationship between the embedded depth of the horizontal drain and the time costs is plotted in Fig. 5. The influence of different phreatic lines (see Table 1) on the position design of the horizontal drain is considered. Compared with the traditional design scheme (Wang *et al.*, 2017; Lei *et al.*, 2016), the horizontal drain at the optimal position can further shorten the consolidation time. A large h_1 will cause the optimal position to move downwards; conversely, the optimal position will move upwards as h_1 decreases. The consolidation rate of the SUSGS with the sand blanket at the optimal position significantly increased (20–80 times) compared to that without the horizontal drain, implying that the horizontal drain shortens the consolidation time and improves engineering efficiency (Zhou *et al.*, 2024).

Table 1. Thickness of unsaturated and saturated soils under different case conditions.

Case	h_1 [m]	h_2 [m]
1	3	7
2	5	5
3	7	3

5.2. The position design of two horizontal drains

If the two horizontal drains are embedded in the SUSGS, the objective function is

$$\min T_v = T_{90}(z_{s1}, z_{s2}). \tag{5.2}$$

Figure 6 shows the distribution of time costs under different depths z_{s1} and z_{s2}, under the condition of $S^* = 90\,\%$. By comparing Figs. 5 and 6, the two horizontal drains further accelerate

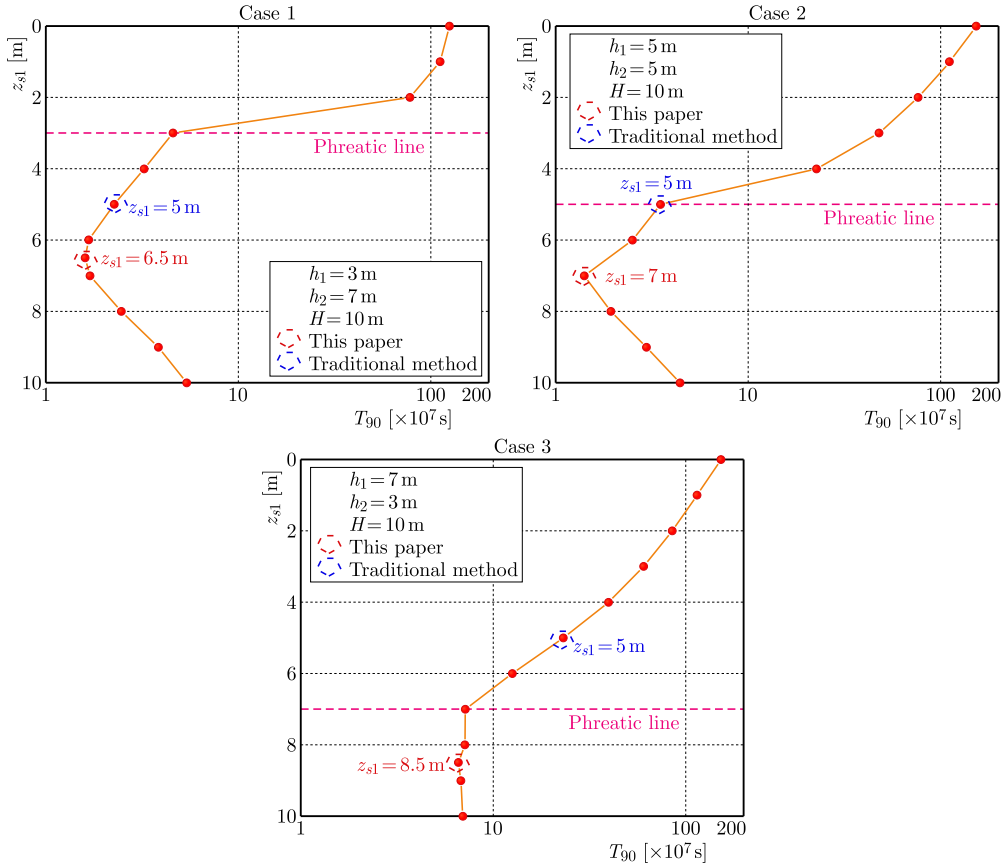

Fig. 5. Relationship between the embedded depth of the horizontal drain and the time costs when $S^* = 90\,\%$.

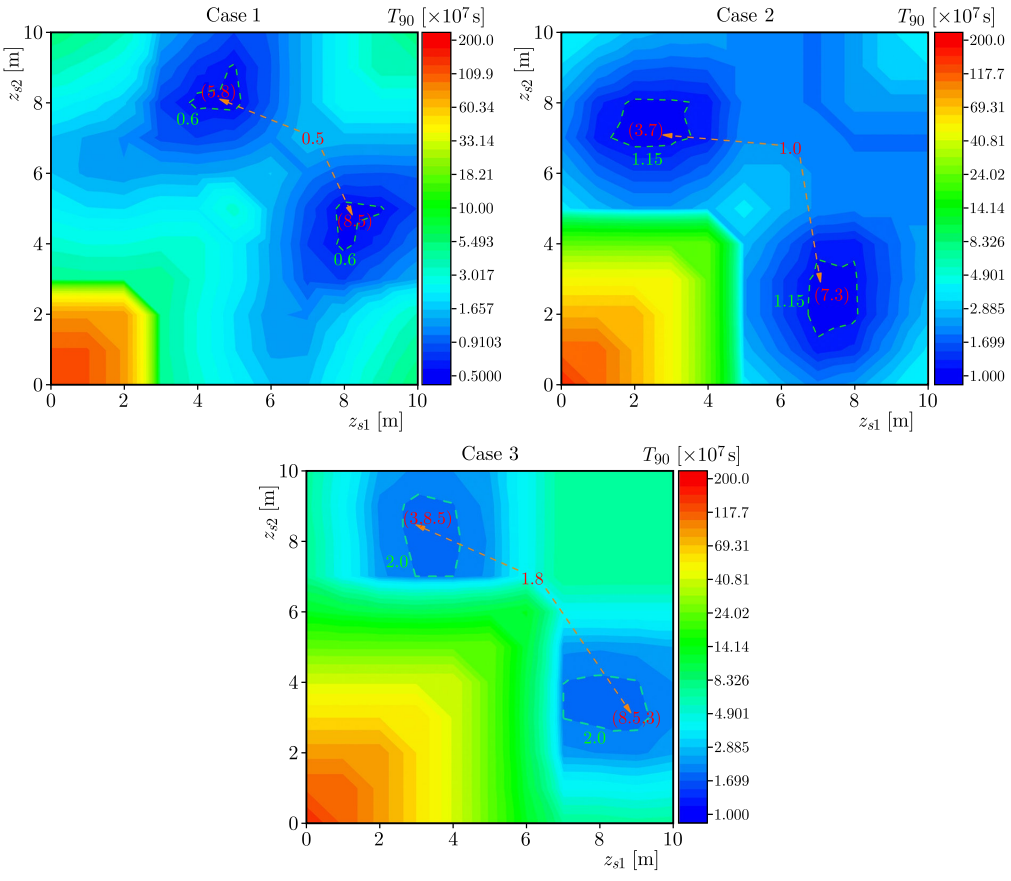

Fig. 6. Time distribution of different depth z_{s1} and z_{s2} when S_{90}.

the consolidation rate. In case 1, the smaller time region (green dashed line) can be found if the horizontal drains are located in the saturated soil layer. The smaller the proportion of unsaturated soil, the more the discharge of excess pore-water pressure in the saturated soil is clogged at the saturated-unsaturated interface. Therefore, there is no need to place a horizontal drain in the unsaturated soil layer. In cases 2 and 3, only when the two horizontal drains are placed in different soil layers, the optimal position of drains in the SUSGS denoted by the smaller time region is obtained. In addition, the larger the proportion of unsaturated soil (i.e., the larger h_1), the slower the consolidation rate of the SUSGS. The consolidation rate of the SUSGS with two horizontal drains is increased 100–150 times compared to that without the horizontal drain. According to the proposed solution's convenience, the optimal position of the horizontal drains in practical engineering can be preliminarily designed.

Table 2 shows the consolidation time considering the traditional design scheme and the proposed optimal position. The horizontal drains are placed at equal intervals in the traditional design scheme. It is found that T_{90} under the traditional design scheme is larger than that under the optimal scheme. If three horizontal drains are located in the SUSGS, the shortest consolidation time is spent. But it may need more engineering resources. A more reasonable choice is to arrange two horizontal drains in the SUSGS. The scheme provided in this paper can achieve a balance between consolidation time and engineering resources.

Table 2. T_{90} (days) for different design schemes for the horizontal drain.

Case	Without horizontal drain	Number of horizontal drains				
		1	2	3	1	2
1	14 467	*262*	*163*	*60*	**189**	**58**
2	17 824	*406*	*176*	*88*	**236**	**115**
3	17 708	*2685*	*1140*	*113*	**775**	**208**

Note: Italicized numbers represent the consolidation time under traditional design schemes. Bold numbers denote the consolidation time under the design scheme of this article.

6. Conclusions

This paper proposes a numerical solution with high computational efficiency and accuracy for designing the position of the horizontal drains in the SUSGS. Based on the consolidation theories of the unsaturated and saturated soils, the two sets of diffusion equations are used to simulate the transient flow of air and water phases. Then, the diffusion equations are discretized by the C-N scheme to obtain the numerical solution for the SUSGS. The proposed numerical solution is reliable and accurate when compared with the analytical solution. According to the verified solution, the influence of the horizontal drains on the consolidation behavior of the SUSGS is investigated to design its optimal position. The results show that the arrangement of the horizontal drain can significantly accelerate the consolidation rate of the entire ground. The horizontal drain at the optimal position can save consolidation time and promote the consolidation speed.

Appendix A

$$\mathbf{A}_u = \begin{pmatrix} 1-\lambda_z c_{vz}^a & c_a & 0.5\lambda_z c_{vz}^a & \cdots & 0 & 0 \\ c_w & 1-\lambda_z c_{vz}^w & 0 & 0.5\lambda_z c_{vz}^w & & \\ 0.5\lambda_z c_{vz}^a & c_w & 1-\lambda_z c_{vz}^a & & & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & 0.5\lambda_z c_{vz}^a & 0 & 1-\lambda_z c_{vz}^a & c_a \\ 0 & 0 & & 0.5\lambda_z c_{vz}^w & c_a & 1-\lambda_z c_{vz}^w \end{pmatrix}, \tag{A.1}$$

$$\mathbf{B}_u = \begin{pmatrix} 1+\lambda_z c_{vz}^a & c_a & -0.5\lambda_z c_{vz}^a & \cdots & 0 & 0 \\ c_w & 1+\lambda_z c_{vz}^w & 0 & -0.5\lambda_z c_{vz}^w & & \\ -0.5\lambda_z c_{vz}^a & c_w & 1+\lambda_z c_{vz}^a & & & \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ & & -0.5\lambda_z c_{vz}^a & 0 & 1+\lambda_z c_{vz}^a & c_a \\ 0 & 0 & & -0.5\lambda_z c_{vz}^w & c_w & 1+\lambda_z c_{vz}^w \end{pmatrix}, \tag{A.2}$$

$$\mathbf{A}_s = \begin{pmatrix} 1-\lambda_z c_{vz} & 0.5\lambda_z c_{vz} & \cdots & 0 & 0 \\ 0.5\lambda_z c_{vz} & 1-\lambda_z c_{vz} & & & \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ & & & 1-\lambda_z c_{vz} & 0.5\lambda_z c_{vz} \\ 0 & 0 & \cdots & 0.5\lambda_z c_{vz} & 1-\lambda_z c_{vz} \end{pmatrix}, \tag{A.3}$$

$$\mathbf{B}_s = \begin{pmatrix} 1+\lambda_z c_{vz} & -0.5\lambda_z c_{vz} & \cdots & 0 & 0 \\ -0.5\lambda_z c_{vz} & 1+\lambda_z c_{vz} & & & \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ & & & 1+\lambda_z c_{vz} & -0.5\lambda_z c_{vz} \\ 0 & 0 & \cdots & -0.5\lambda_z c_{vz} & 1+\lambda_z c_{vz} \end{pmatrix}, \tag{A.4}$$

$$\mathbf{D}_u^n = \begin{bmatrix} u_{a1}^{(1)n} & u_{w1}^{(1)n} & \cdots & u_{a(M-1)}^{(1)n} & u_{w(M-1)}^{(1)n} \end{bmatrix}^{\mathrm{T}}, \tag{A.5}$$

$$\mathbf{D}_s^n = \begin{bmatrix} u_{w(M+1)}^{(2)n} & u_{w(M+2)}^{(2)n} & \cdots & u_{w(J-2)}^{(2)n} & u_{w(J-1)}^{(2)n} \end{bmatrix}^{\mathrm{T}}, \tag{A.6}$$

$$\mathbf{Q}_u^n = \left(q^{n+1} - q^n\right) \begin{bmatrix} c_\sigma^a & c_\sigma^w & \cdots & c_\sigma^a & c_\sigma^w \end{bmatrix}^{\mathrm{T}}, \tag{A.7}$$

$$\mathbf{Q}_s^n = \left(q^{n+1} - q^n\right) \begin{bmatrix} 1 & 1 & \cdots & 1 & 1 \end{bmatrix}^{\mathrm{T}}, \tag{A.8}$$

where $\mathbf{A}_u$ and $\mathbf{B}_u$ are five-diagonal sparse matrices, $\mathbf{A}_s$ and $\mathbf{B}_s$ are tridiagonal sparse matrices.

Appendix B

Derivation of stability

In the von Neumann criterion, a standardized trial solution is proposed to verify the stability of the difference scheme. The trial solution of Eq. (3.3) at point (z_k, t_n) is assumed to be:

$$\mathbf{u}_j^n = \mathbf{v}^n e^{ik\zeta h_z}, \tag{B.1}$$

where $\mathbf{u}_k^n = [u_{ak}^n \;\; u_{wk}^n]^{\mathrm{T}}$, $i^2 = -1$, ζ reflects the small errors caused by node j iterating in the adjacent time layers.

Substituting Eq. (B.1) into Eq. (3.3) yields:

$$\mathbf{G}_1 \mathbf{v}^{n+1} = \mathbf{G}_2 \mathbf{v}^n, \tag{B.2}$$

where

$$\mathbf{G}_{u1}=\begin{pmatrix}1-2\lambda_z\alpha c_{vz}^{a} & c_a\\ c_w & 1-2\lambda_z\alpha c_{vz}^{w}\end{pmatrix},\qquad \mathbf{G}_{u2}=\begin{pmatrix}1+2\lambda_z\alpha c_{vz}^{a} & c_a\\ c_w & 1+2\lambda_z\alpha c_{vz}^{w}\end{pmatrix},$$

$$\alpha=\left(\sin\left(\frac{\zeta h}{2}\right)\right)^2.$$

The growth matrix of Eq. (3.3) is

$$\mathbf{G}_u=(\mathbf{G}_{u1})^{-1}\,\mathbf{G}_{u2}. \tag{B.3}$$

Based on Eq. (B.3), the eigenvalues of the growth matrix are

$$\mu_1=\frac{c_ac_w+4a^2\lambda_z^2c_{vz}^ac_{vz}^w-1-2a\lambda_z\left[(c_{vz}^a)^2-2c_{vz}^ac_{vz}^w+(c_{vz}^w)^2+4c_ac_wc_{vz}^ac_{vz}^w\right]^{1/2}}{c_ac_w-4a^2\lambda_z^2c_{vz}^ac_{vz}^w-1+2a\lambda_zc_{vz}^a+2a\lambda_zc_{vz}^w}, \tag{B.4}$$

$$\mu_2=\frac{c_ac_w+4a^2\lambda_z^2c_{vz}^ac_{vz}^w-1+2a\lambda_z\left[(c_{vz}^a)^2-2c_{vz}^ac_{vz}^w+(c_{vz}^w)^2+4c_ac_wc_{vz}^ac_{vz}^w\right]^{1/2}}{c_ac_w-4a^2\lambda_z^2c_{vz}^ac_{vz}^w-1+2a\lambda_zc_{vz}^a+2a\lambda_zc_{vz}^w}. \tag{B.5}$$

According to the von Neumann criterion, the difference scheme is unconditionally stable only when all absolute values of eigenvalues ($|u_1 \sim u_2| \leq 1$) are less than or equal to 1. Otherwise, it is conditionally stable. Based on Eqs. (B.4) and (B.5), considering that the consolidation parameters (c_{vz}^a and c_{vz}^w) of soil are very small, the denominator is always greater than the numerator in the eigenvalues. Thus, the C-N scheme for the consolidation equations of the unsaturated soil is unconditionally stable under the arbitrary mesh ratio (λ_z).

Similarly, the trial solution of Eq. (3.4) at point (z_k,t_n) is assumed to be:

$$u_{wj}^{(2)n}=v_w^{(2)n}e^{ik\zeta h}. \tag{B.6}$$

Substituting Eq. (B.6) into Eq. (3.4) yields:

$$G_{s1}v_w^{(2)n+1}=G_{s2}v_w^{(2)n}, \tag{B.7}$$

where $G_{s1}=1-2\lambda_z\alpha c_{vz}$, $G_{s2}=1+2\lambda_z\alpha c_{vz}$.

The growth factor of Eq. (3.4) is:

$$G_s=G_{s2}/G_{s1}=\frac{1+2\lambda_z\alpha c_{vz}}{1-2\lambda_z\alpha c_{vz}}. \tag{B.8}$$

It can be found that growth factor G_s is less than 1 under the arbitrary mesh ratio, so the C-N solution for the consolidation equation of the saturated soil is also unconditionally stable.

Appendix C

A and **B** are the partitioned matrices, and they are expressed as

$$\mathbf{A}=\begin{pmatrix}\mathbf{A}_0 & \cdots & \mathbf{0} & \cdots & \mathbf{0}\\ & \ddots & & & \vdots\\ \vdots & & \ddots & & \mathbf{0}\\ \mathbf{0} & & & \ddots & \vdots\\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \mathbf{A}_J\end{pmatrix}\quad \begin{matrix}\mathbf{A}_0,\ \mathbf{A}_J\\ \mathbf{A}_u\\ \mathbf{A}_M\\ \mathbf{A}_s\\ \mathbf{A}_{k1},\ \mathbf{A}_{k2}\end{matrix} \tag{C.1}$$

where $\mathbf{A}_0$ and $\mathbf{A}_J$ are the coefficient matrices of Eqs. (3.3) and (3.4) at the boundary $z = z_0$ and $z = z_J$, respectively, $\mathbf{A}_M$ is coefficient matrices on the left side of Eq. (3.13), $\mathbf{A}_{m1}$ and $\mathbf{A}_{m2}$ are identity matrices:

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_0,\ \mathbf{B}_J & \cdots & \mathbf{0} & \cdots & \mathbf{0} \\ & \mathbf{B}_u & & & \mathbf{0} \\ \vdots & & \mathbf{B}_M & & \vdots \\ \mathbf{0} & & & \mathbf{B}_s & \mathbf{0} \\ \vdots & & & & \mathbf{B}_{k1},\ \mathbf{B}_{k2} \\ \mathbf{0} & \cdots & \mathbf{0} & \cdots & \end{pmatrix} \tag{C.2}$$

Appendix D

Comparison of computational efficiency and accuracy

As shown in Tables D1 and D2, the numerical solution developed in this paper has high computational efficiency and stability under the arbitrary mesh ratio condition. It is worth noting that the C-N scheme at the local region has been proven to be unconditionally stable. In order to verify the overall stability of the C-N solution, the boundary conditions, initial conditions, and external loading are considered in this example. It can be seen that the numerical solution also exhibits strong stability, indicating that the C-N solution can solve consolidation problems under various boundary conditions, initial conditions, and time-dependent loading.

Table D1. Comparison of the calculation results between the numerical and analytical solutions.

Time step τ [s]	Time costs [s]	Water pressure [kPa], $h_z = 1\,\mathrm{m}$, $t_n = 10^7\,\mathrm{s}$			
		$z = 1\,\mathrm{m}$	$z = 3\,\mathrm{m}$	$z = 5\,\mathrm{m}$	$z = 7\,\mathrm{m}$
10	258.25	20.33	32.64	86.04	90.13
100	24.32	20.33	32.64	86.04	90.13
1000	2.26	20.33	32.64	86.04	90.13
10 000	0.35	20.33	32.64	86.04	90.13
Yuan *et al.* (2023)	–	20.34	32.64	86.05	90.13

Note: $t_N = 10^9\,\mathrm{s}$

Table D2. Comparison of the calculation results between the numerical and analytical solutions.

Depth step h_z	Water pressure [kPa], $\tau = 100\,\mathrm{s}$, $t_n = 10^7\,\mathrm{s}$			
	$z = 1\,\mathrm{m}$	$z = 3\,\mathrm{m}$	$z = 5\,\mathrm{m}$	$z = 7\,\mathrm{m}$
0.1	20.33	32.64	86.04	90.13
0.2	20.33	32.64	86.04	90.13
0.25	20.33	32.64	86.04	90.13
0.5	20.33	32.64	86.04	90.13
1	20.33	32.64	86.04	90.13

Acknowledgments

The authors gratefully acknowledge the financial support for this work from the National Engineering Laboratory for Highway Tunnel Construction Technology (grant no. NELFHT201702), Natural Science Foundation of Sichuan (grant no. 2022NSFSC0443).

References

1. Feng, J.X., Ni, P.P., Chen, Z., Mei, G.X., & Xu, M.J. (2020). Positioning design of horizontal drain in sandwiched clay-drain systems for land reclamation. *Computers and Geotechnics*, *127*, Article 103777. https://doi.org/10.1016/j.compgeo.2020.103777

2. Feng, J.X., Ni, P.P., & Mei, G.X. (2019). One-dimensional self-weight consolidation with continuous drainage boundary conditions: Solution and application to clay-drain reclamation. *International Journal for Numerical and Analytical Methods in Geomechanics*, *43*(8), 1634–1652. https://doi.org/10.1002/nag.2928

3. Fredlund, D.G., & Hasan, J.U. (1979). One-dimensional consolidation theory: unsaturated soils. *Canadian Geotechnical Journal*, *16*(3), 521–531. https://doi.org/10.1139/t79-058

4. Fredlund, D.G., Rahardjo, H., & Fredlund, M.D. (2012). *Unsaturated soil mechanics in engineering practice*. John Wiley & Sons, Inc. https://doi.org/10.1002/9781118280492

5. Gibson, R.E., & Shefford, G.C. (1968). The efficiency of horizontal drainage layers for accelerating consolidation of clay embankments. *Géotechnique*, *18*(3), 327–335. https://doi.org/10.1680/geot.1968.18.3.327

6. Gu, L.L., Wang, Z., Huang, Q., Ye, G.L., & Zhang, F. (2020). Numerical investigation into ground treatment to mitigate the permanent train-induced deformation of pile-raft-soft soil system. *Transportation Geotechnics*, *24*, Article 100368. https://doi.org/10.1016/j.trgeo.2020.100368

7. He, Q., Yuan, Q., Lang, L., & Cao, J.X. (2025). Implicit difference solution for one-dimensional consolidation of unsaturated soil: Numerical verification and application to soil-cushion system. *Mathematical Geosciences*, *57*, 951–976. https://doi.org/10.1007/s11004-025-10177-6

8. Lee, S.L., Karunaratne, G.P., Yong, K.Y., & Ganeshan, V. (1987). Layered clay-sand scheme of land reclamation. *Journal of Geotechnical Engineering*, *113*(9), 984–995. https://doi.org/10.1061/(ASCE)0733-9410(1987)113:9(984)

9. Lei, G.H., Li, Z., & Xu, L.D. (2016). Free-strain solutions for two-dimensional consolidation with sand blankets (in Chinese). *Chinese Journal of Geotechnical Engineering*, *38*(2), 193–201. https://doi.org/10.11779/CJGE201602001

10. Li, H.P., Chen, Z., Feng, J.X., Meng, Y.H., & Mei, G.X. (2020). Study on position optimization of horizontal drainage sand blanket of double-layer foundation (in Chinese). *Rock and Soil Mechanics*, *41*(2), 437–444.

11. Li, L.Z., Qin, A.F., & Jiang, L.H. (2021). Semi-analytical solution for one-dimensional consolidation of a two-layered soil system with unsaturated and saturated conditions. *International Journal for Numerical and Analytical Methods in Geomechanics*, *45*(15), 2284–2300. https://doi.org/10.1002/nag.3266

12. Li, L.Z., Qin, A.F., Jiang, L.H., & Wang, L. (2022). One-dimensional consolidation of unsaturated–saturated soil system considering pervious or impervious drainage condition induced by time-dependent loading. *Computers and Geotechnics*, *152*, Article 105053. https://doi.org/10.1016/j.compgeo.2022.105053

13. Meng, Y.H., Chen, Z., Feng, J.X., Li, H.P., & Mei, G.X. (2019). Optimization of one-dimensional foundation with sand blankets under the non-uniform distribution of initial excess pore water pressure (in Chinese). *Rock and Soil Mechanics*, *40*(12), 4793–4800. http://doi.org/10.16285/j.rsm.2018.1899

14. Mesri, G., & Funk, J.R. (2015). Settlement of the Kansai International Airport Islands. *Journal of Geotechnical and Geoenvironmental Engineering*, *141*(2), Article 04014102. https://doi.org/10.1061/(ASCE)GT.1943-5606.0001224

15. Moradi, M., Keshavarz, A., & Fazeli, A. (2019). One dimensional consolidation of multi-layered unsaturated soil under partially permeable boundary conditions and time-dependent loading. *Computers and Geotechnics*, *107*, 45–54. https://doi.org/10.1016/j.compgeo.2018.11.020

16. Morton, K.W., & Mayers, D.F. (2005). *Numerical solution of partial differential equations: An introduction* (2nd ed.). Cambridge University Press, Cambridge, UK.

17. Qin, A.F., Chen, G.J., Tan, Y.W., & Sun, D.A. (2008). Analytical solution to one-dimensional consolidation in unsaturated soils. *Applied Mathematics and Mechanics – English Edition*, *29*, 1329–1340. https://doi.org/10.1007/s10483-008-1008-x

18. Sharifi, S., & Rashidinia, J. (2016). Numerical solution of hyperbolic telegraph equation by cubic B-spline collocation method. *Applied Mathematics and Computation*, *281*, 28–38. https://doi.org/10.1016/j.amc.2016.01.049

19. Terzaghi, K. (1943). *Theoretical soil mechanics*. John Wiley and Sons, Inc. https://doi.org/10.1002/9780470172766

20. Wang, L., Sun, D.A., Li, L.H., Li, P.C., & Xu, Y.F. (2017). Semi-analytical solutions to one-dimensional consolidation for unsaturated soils with symmetric semi-permeable drainage boundary. *Computers and Geotechnics*, *89*, 71–80. https://doi.org/10.1016/j.compgeo.2017.04.005

21. Wang, L., Xu, Y.F., Xia, X.H., He, Y.L., & Li, T.Y. (2019). Semi-analytical solutions to two-dimensional plane strain consolidation for unsaturated soils under time-dependent loading. *Computers and Geotechnics*, *109*, 144–165. https://doi.org/10.1016/j.compgeo.2019.01.002

22. Wu, J., Xi, R., Liang, R., Zong, M., & Wu, W. (2023). One-dimensional nonlinear consolidation for soft clays with continuous drainage boundary considering non-Darcy flow. *Applied Sciences*, *13*(6), Article 3724. https://doi.org/10.3390/app13063724

23. Yuan, Q., He, Q., Lang, L., & Yang, X.B. (2023). Analytical solution for Fredlund-Hasan unsaturated consolidation using mode superposition method. *International Journal for Numerical and Analytical Methods in Geomechanics*, *47*(17), 3234–3247. https://doi.org/10.1002/nag.3618

24. Zhou, T., Wang, L., Li, T.Y., Wen, M.J., & Zhou, A.N. (2023). Semi-analytical solutions to the one-dimensional consolidation for double-layered unsaturated ground with a horizontal drainage layer. *Transportation Geotechnics*, *38*, Article 100909. https://doi.org/10.1016/j.trgeo.2022.100909

25. Zhou, T., Yan, X.L., Wang, L., Zhou, A.N., & Sun, D.A. (2024). Semi-analytical solution for one-dimensional consolidation of a two-layered unsaturated-saturated soil system ground with a horizontal drainage layer. *Computers and Geotechnics*, *169*, Article 106196. https://doi.org/10.1016/j.compgeo.2024.106196

Manuscript received September 20, 2024; accepted for publication April 3, 2025; published online September 27, 2025.

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 855–864, Warsaw 2025
https://doi.org/10.15632/jtam-pl/204537

THE IMPACT OF RACKET STRING TENSION ON TENNIS RACKET PERFORMANCE

Mingshun JIA, Xiaobin LAN, Tao ZHUANG, Bo SUN*

School of Physical Education, Liaocheng University, Liaocheng, China
*corresponding author, *sunbo@lcu.edu.cn*

This paper investigates the impact of tennis racket string tension, within the same material, on the effectiveness of ball striking. The images were recorded by a high-speed camera, and 8 kinds of string pounds were used in the experiment to fix the frame of the tennis racket, and a homemade ball dispenser was used to launch tennis balls with different incidence velocities and angles to the geometric center of the racket head. Within the tested range, the 40-pound string tension demonstrated the highest coefficient of restitution (COR), while the 70-pound tension showed a comparatively lower value.

Keywords: string pounds; coefficient of restitution; angle of rebound; contact time.

1. Introduction

In the sport of tennis, racket string tension is one of the key factors affecting racket performance. It not only determines the elasticity and energy transfer efficiency of the racket but also directly influences the player's control over the ball and the ball's striking speed (Ghaednia *et al.*, 2015). Research on speed aspects is commonly evaluated using the coefficient of restitution (COR), which represents the ratio of the ball's rebound speed to its pre-impact speed (Cross, 1999). Variations in the COR reflect the elastic characteristics of the interaction between the ball and the racket strings, which are influenced by the string tension (Cross & Bower, 2001). Lower string tensions can produce faster rebound speeds, primarily because the string bed deforms more during impact, the ball deforms less, and the strings can store more energy to return to the tennis ball as kinetic energy, causing the ball to rebound at a faster speed (Cross, 2000; Haake *et al.*, 2003; Kotze *et al.*, 2000). Existing studies have indicated that the selection of string tension should not be too low, nylon strings at around 50 pounds of tension can rebound at faster speeds (Baker & Wilson, 1978; Bower & Cross, 2008; Goodwill & Haake, 2004; Hatch *et al.*, 2006).

String tension and incident speed are key factors affecting the rebound angle (Bower & Cross, 2005; Bower & Sinclair, 1999), and for the same angle of incidence, the smaller the rebound angle, the easier it is for the racket to control the ball, and the better the stability of the racket. The rebound angle of a tennis ball is an important indicator for testing the stability of a racket's control over the ball. Goodwill and Haake (2004) conducted simulated ball-striking experiments with rackets at 40 pounds and 70 pounds of string tension and found that the rebound angle of the tennis ball was influenced by the string tension, with the racket at 40 pounds showing a lower rebound angle than the 70-pound racket. Bower and Sinclair (1999) used rackets with string tensions of 40 pounds, 51 pounds, and 62 pounds for their experiments and found that when the tennis ball struck the 40-pound racket, the rebound angle was closer to the normal of the racket face, while the 62-pound racket produced a rebound angle further away from the normal of the racket face.

In the sport of tennis, the interaction between the ball and the racket has a decisive impact on the outcome of matches (Cross *et al.*, 2000). Particularly, the string tension of the racket and the contact time between the ball and the string are key parameters affecting racket performance and player performance (Bao *et al.*, 2003). The tension of the strings not only affects the rotation and controllability of the ball but is also directly related to the player's feel and comfort (Bendtsen *et al.*, 2015; Bower & Cross, 2003; Hennig, 2007; Kawazoe *et al.*, 2012). In actual competition, the contact time between the ball and the strings is extremely short, usually at the millisecond level (Miller, 2005). This brief contact time significantly influences the ball's speed, spin, and trajectory (Bao *et al.*, 2003; Cross, 2003). Goodwill and Haake (2004) found that the contact time for a 40-pound string is higher than that for a 70-pound racket, suggesting a negative correlation between string tension and contact time. Baktiar *et al.* (2020) conducted impact force tests on two rackets with different tensions and the results indicated that the higher the string tension, the shorter the duration of the collision.

For most recreational players, opting for lower string tension (e.g., 48 lbs) can enhance the racket's elasticity, making it easier to hit the ball and improving both the striking success rate and speed. Medium string tension (e.g., 54 lbs) offers better striking effect and control, while reducing the impact on the forearm (Zhao *et al.*, 2025). Excessively high tension may lead to poor ball placement control and increased errors, whereas overly low tension, though boosting ball speed, can compromise accuracy (Baszczyński *et al.*, 2016). Professional players choose string tension based on match specifics and their technical strengths. For instance, they may select higher tension for more spin and control on certain courts and lower tension for greater speed and coverage on others. When choosing string tension, professionals prioritize the ball's rebound characteristics and the racket's control across various strokes (Allen *et al.*, 2016; Chadefaux *et al.*, 2016).

The current research has the following limitations. First, previous studies have mainly focused on nylon strings, and their findings may not directly apply to other common materials like polyester fiber strings, which have different properties affecting the ball-hitting effect. Second, prior research has a relatively narrow scope regarding racket string tension, impact speed, and incident angle. To better understand the complex interaction between tennis balls and racket strings, these variables need wider-ranging research. Third, while some studies have explored the impact of racket string tension on specific parameters (e.g., coefficient of restitution or rebound angle), they lack comprehensive analysis integrating multiple factors, thus failing to fully understand how string tension affects racket performance. This study selects commonly-used polyester strings, experiments with various tensions, three impact speeds, and five incident angles, and compares the results with previous studies. By studying the effect of different racket string tensions on the rebound speed and angle of tennis balls after impact, it can be found that the same incident speed but different string tensions leads to different rebound speeds. A faster rebound speed indicates a better striking effect, enabling athletes to generate higher ball speeds. A smaller rebound angle means less striking deviation and better ball control, making it easier for athletes to direct the ball to the desired position. Conversely, a larger rebound angle results in poorer ball control, increasing the risk of the ball going out of bounds or having a larger landing point deviation.

2. Materials and methods

2.1. Experimental materials

Wilson Tennis Racket (Wilson Pro Staff v10): The unstrung racket weighs 315 g and is strung with 16 main strings and 19 cross strings. The strung racket weighs 340 g, has a head size of 626 cm^2 (97 square inches), and the racket length is 68.58 cm (27 inches). Luxilon tennis strings (35, 40, 45, 50, 55, 60, 65, 70 pounds): Made of polyester material, with a string diameter

of 1.25 mm. Slazenger Tennis Balls: The balls weigh 58 g each. Two light sources, a reflector, a tripod, and a suitably sized wooden board are used to secure the racket; the central part of the board corresponding to the racket head is hollowed out to prevent the string bed from contacting the board upon impact at the racket head.

2.2. Experimental equipment

Before the experiment, a vise and an angle meter were used to adjust the desired angle. A high-speed camera (NX3-S3) produced by IDT Corporation in the United States, with a shooting frequency of 1500 frames per second and a lens focal length of 16 mm, was employed. The scale bar is 0.23 m in the horizontal direction and 0.5 m in the vertical direction. Finally, SIMI-motion software from Germany was used for image analysis.

2.3. Experimental scheme

Before the experiment, a marker was used to label the tennis ball for better determination of the geometric center of the racket face, and the reflective tape was applied for marking. The main optical axis of the high-speed camera was aligned at the same height as the center of the racket head, with the camera position fixed. The tennis ball was placed at the geometric center of the racket, and manual aperture adjustment was made for focusing, with the ability to accurately recognize the markings on the tennis ball as a reference. A plane mirror was positioned at a 45-degree angle in front of the side of the racket to ascertain whether the tennis ball struck the geometric center of the racket face. The ball-feeding machine was adjusted and placed approximately 1.5 m away from the racket, with the prepared scale bar placed at the geometric center of the racket for filming (Fig. 1). In this study, eight racket string tensions were tested. For the 35 lbs tension, impacts at 0°, 30°, 40°, 45°, and 50° incident angles were performed at 20 m/s, 25 m/s, and 30 m/s using a homemade launcher aimed at the racket's geometric center. After the 35 lbs tests, the strings were cut, and a stringing machine was used to restring the racket for the 40 lbs tests, and so on for subsequent tensions.

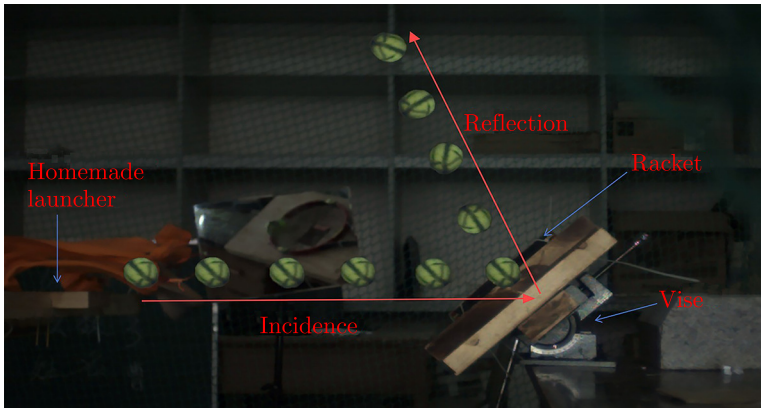

Fig. 1. Superimposition of multiple images of actual tennis incidence and rebound.

2.4. Data processing

The data were statistically processed using a Microsoft Excel spreadsheet for each beat-string poundage, plotted using origin2021 software to plot the selected parameters as a scatterplot, and fitted to the data based on the scatterplot trend. If close to linear, a linear fit was used. In this experiment, the angle between the incident line and the normal is called the angle of incidence α, and the angle between the reflected line and the normal is called the angle of reflection β. The partial velocity perpendicular to the y-axis direction of the racket surface is denoted by

$$v_{yi} = v_i \cdot \cos\alpha. \tag{2.1}$$

The partial velocity of the tennis ball after rebound is expressed as

$$v_{yr} = v_r \cdot \cos\beta. \tag{2.2}$$

Calculation of the COR is expressed as

$$e = \frac{v_{yr}}{v_{yi}}. \tag{2.3}$$

The ratio of the angle of rebound to the angle of incidence is expressed as

$$\delta = \frac{\beta}{\alpha}. \tag{2.4}$$

Select two fixed points at the top and bottom of the racket head to establish a straight line connecting these two points. Identify the frame immediately preceding the moment of ball impact as the incidence moment, and use the center of the ball in this frame as a fixed point. Connect the centers of the ball from all frames before the impact with this fixed point to obtain multiple straight lines representing the ball's trajectory before impact. For the frame in which the ball leaves the racket after impact, consider it as the bounce moment, and use the center of the ball in this frame as a fixed point. Connect the centers of the ball from all frames following the impact with this fixed point to obtain multiple straight lines representing the ball's trajectory after impact. By connecting the obtained straight lines before the ball impacts with the straight line through the top and bottom points of the racket head, the angles of incidence before the ball strikes the racket can be determined. Calculate the angles of incidence θ by subtracting the complementary angles from 90°. Select the average of the angles of incidence from the 5 consecutive frames with the closest values within the last 10 frames before impact as the incidence angle. Similarly, the rebound angle can be obtained (Fig. 2).

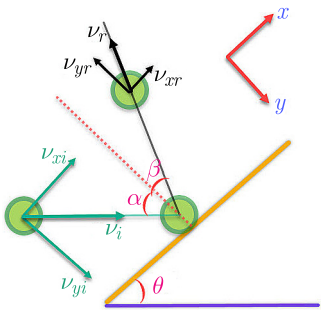

Fig. 2. Schematic diagram of tennis ball incidence and rebound.

3. Results

3.1. Correlation between racket string tension and coefficient of restitution and rebound speed

A fitting equation was established between racket string tension and the COR. At various incidence angles and different ball speeds, there is a correlation between string tension and the COR. The COR tends to decrease with an increase in string tension (Fig. 3). At an incidence angle of 50°, the COR for strings is greater than that at 45°, followed by 40°, then 30°, and the coefficient is the smallest at 0° incidence angle. Moreover, the COR increases with the increase in incidence angle. At 0° incidence angle, the COR decreases with the increase in incidence speed. At an incidence speed of 20 m/s, the coefficients of restitution for 40-pound and 35-pound strings are close and greater than those of other tensions, while the coefficient for 70-pound strings is smaller (Fig. 3a). At incidence speeds of 25 m/s and 30 m/s, the COR for

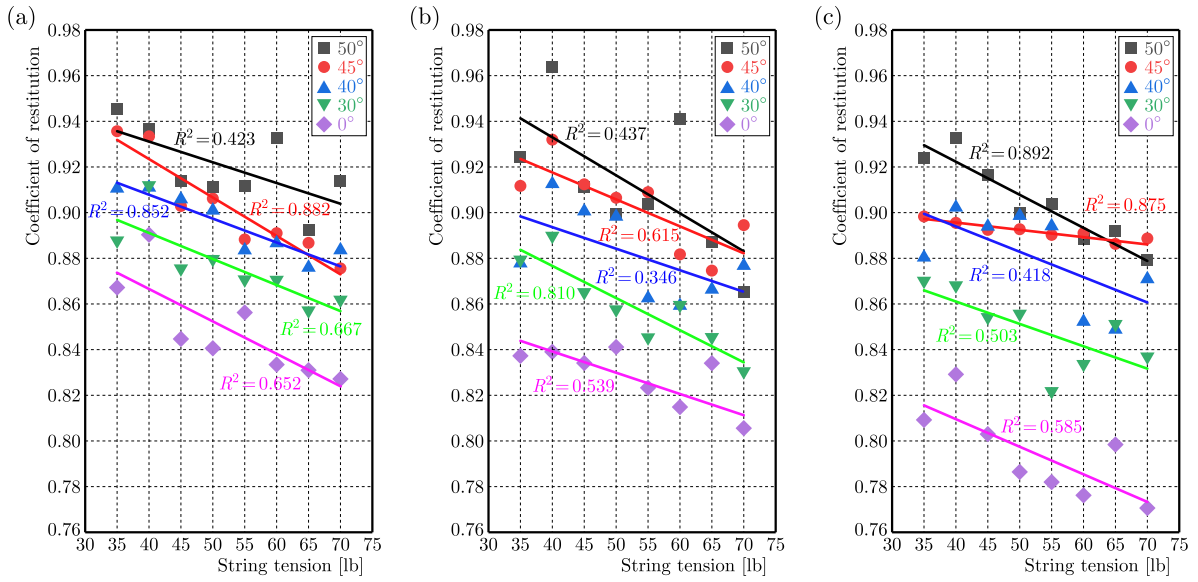

Fig. 3. Linear fitting plots of racket string tension versus coefficient of restitution at various incidence angles and ball speeds: (a) incident speed of 20 m/s; (b) incident speed of 25 m/s; (c) incident speed of 30 m/s.

40-pound strings is greater than that for other tensions, and the coefficient for 70-pound strings is smaller (Fig. 3b and 3c). Additionally, it can be observed from the figures that the COR decreases as the string tension is reduced from 40 pounds to 35 pounds.

By establishing a fitting equation between string tension and rebound speed, it is found that under various incidence angles and ball speeds, the rebound speed decreases as the string tension increases. Specifically, the rebound speeds at 40 and 35 pounds of string tension are greater than those at other tensions (Fig. 4).

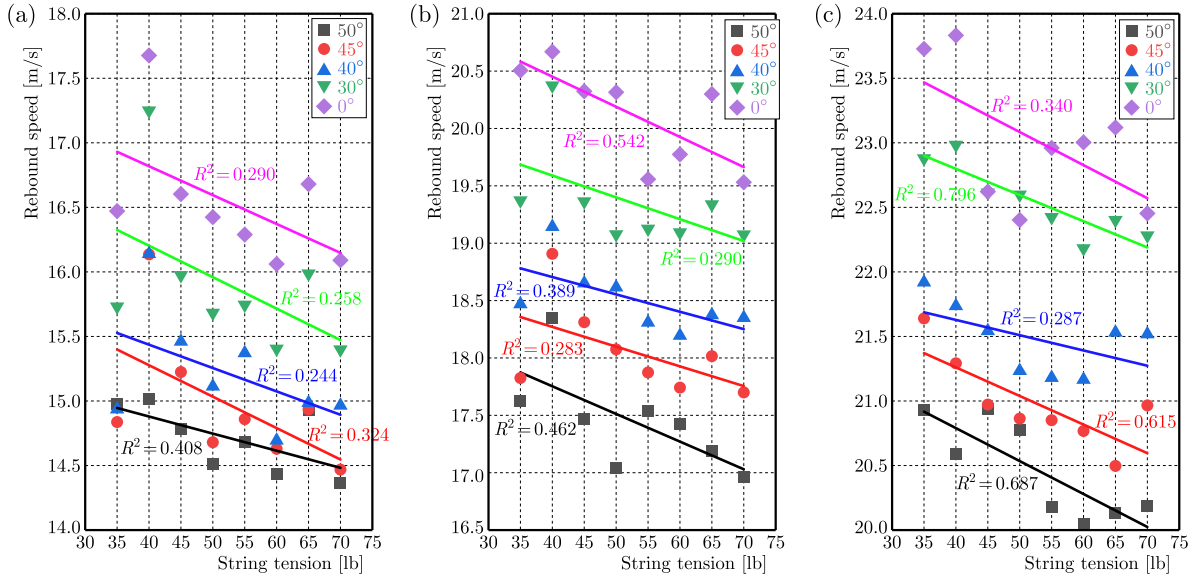

Fig. 4. Linear fitting plots of racket string tension versus rebound speed at various incidence angles and ball speeds: (a) incident speed of 20 m/s; (b) incident speed of 25 m/s; (c) incident speed of 30 m/s.

3.2. Correlation between racket string tension and angle ratio

By establishing a linear fitting equation between racket string tension and the angle ratio, the study results indicate a linear correlation between string tension and the angle ratio. As the

string tension increases, the angle ratio tends to increase (Fig. 5). Furthermore, with the increase in incidence velocity, the angle ratio exhibits an overall decreasing trend. At various incidence angles (excluding 0°), the angle ratio for 35-pound string tension is significantly lower than that for 70-pound string tension. Additionally, at the same string tension, the angle ratio increases with the increase in incidence angle. The angle ratio for strings at an incidence angle of 50° is greater than that at 45°, followed by 40°, with the angle ratio at an incidence angle of 30° being the smallest.

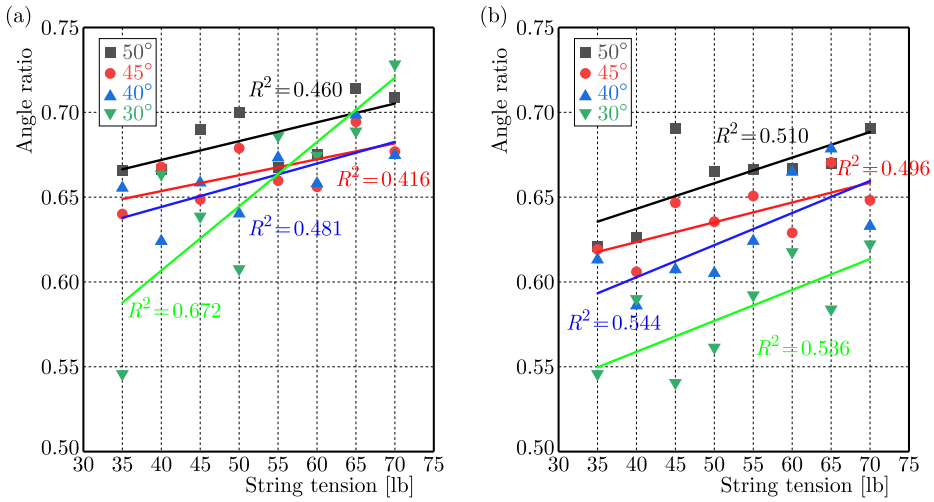

Fig. 5. Linear fitting plots of racket string tension versus angle ratio at various incidence angles and ball speeds: (a) incident speed of 25 m/s; (b) incident speed of 30 m/s.

3.3. Correlation between racket string tension and contact time

By establishing a linear fitting equation between racket string tension and contact time, the study results indicate a linear correlation between the two variables. As the string tension increases, the contact time exhibits a decreasing trend (Fig. 6). Across various incidence angles, the

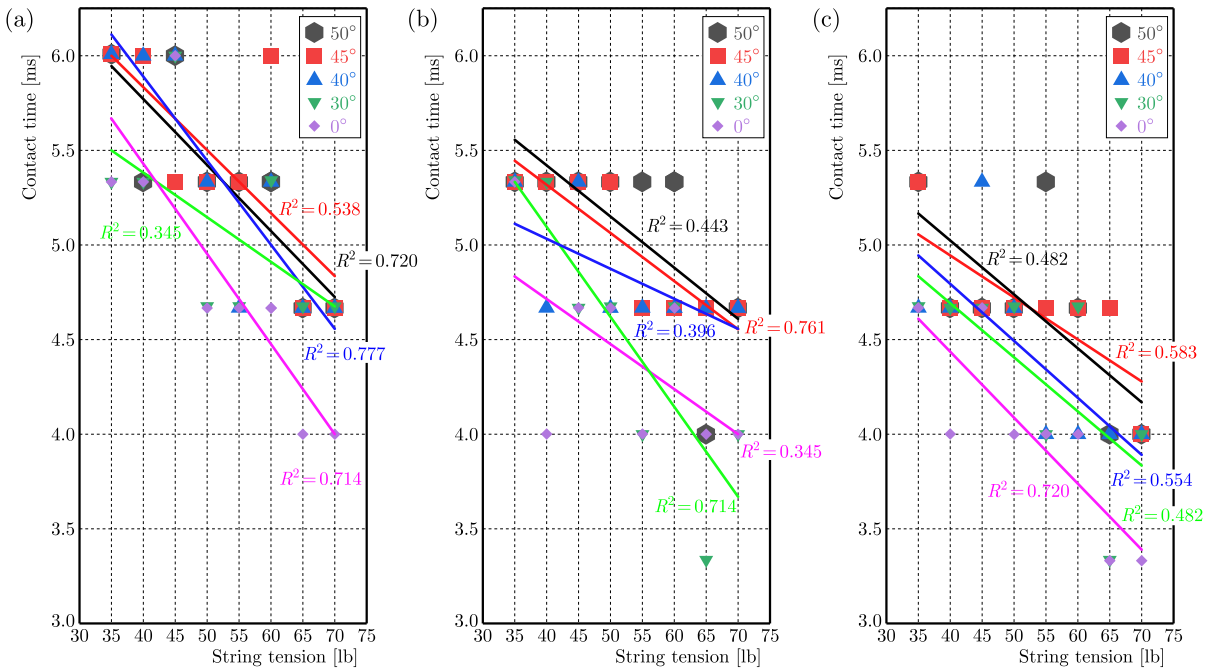

Fig. 6. Linear fitting plots of racket string tension versus contact time at various incidence angles and ball speeds: (a) incident speed of 20 m/s; (b) incident speed of 25 m/s; (c) incident speed of 30 m/s.

contact time for 35-pound tension is longer compared to other tensions, with 70-pound tension having the shortest contact time. At incidence angles of 0° and 30°, the contact time between the tennis ball and the racket strings is shorter than at 40°, 45°, and 50°, suggesting a slight increase in contact time with the amplification of the incidence angle. Across different incidence angles, the contact times in Fig. 6a are overall longer than those in Figs. 6b and 6c, where the contact times are the shortest, demonstrating that with the increase in incidence velocity, the contact time shows a decreasing trend.

4. Discussion

The aforementioned research results indicate that with the increase in incidence velocity, the COR exhibits a decreasing trend (Fig. 3). The possible reason for this is that in the range tested, most of the kinetic energy is dissipated due to the movement inside the racket frame or due to the greater deformation of the tennis ball with the string, resulting in more energy dissipation (Bao *et al.*, 2003). The study found that with the increase in the angle of incidence, the COR tends to increase, which is consistent with the research findings of (Cross, 1999). Additionally, the study discovered that within the tested range, the COR decreases with the increase in racket string tension; the 40-pound string tension exhibits the highest COR, enabling a faster rebound speed for the tennis ball, while the 70-pound string tension shows the lowest COR, resulting in the slowest rebound speed (Fig. 4). The reason may be that at lower string tensions, the string bed can undergo greater deformation, absorbing more energy, which leads to less energy loss for the tennis ball. Consequently, the overall energy loss in the system of the racket and the tennis ball is reduced. Therefore, when the tennis ball leaves the string bed, the strings can transfer a larger portion of energy back to the tennis ball, enabling it to have a faster speed and thus increasing the COR (Baktiar *et al.*, 2020). When the string tension is high, the degree of deformation of the strings is minimal, the whole system energy loss between the string and the tennis ball increases, and the string can only provide a small portion of the energy to the tennis ball as it leaves the string, resulting in a slower tennis ball after rebound, leading to increased energy loss in the system comprising the strings and the tennis ball. Consequently, lower tension strings can enable a faster rebound speed (Bower & Cross, 2005; Cross, 2000; Haake *et al.*, 2003). Within the tested range, a decreasing trend in the COR was observed when the tension was reduced from 40 pounds to 35 pounds. The reason might be that when the tennis ball impacts strings with excessively low tension, the greater degree of string deformation leads to increased energy loss due to excessive movement, resulting in a decrease in the speed of the ball and a decrease in the COR after rebound (Baker & Wilson, 1978).

At various incidence angles (except for 0°), the study indicates that the angle ratio increases with the increase in racket string tension (Fig. 5), and lower tension rackets result in smaller rebound angles, which is consistent with the findings of (Bower & Cross, 2005; Goodwill & Haake, 2004). This suggests that a tennis ball striking a high-tension string (70 pounds) rebounds on a trajectory further away from the racket's normal, which is less favorable for ball control. Conversely, a tennis ball striking a low-tension string (35 pounds) rebounds on a trajectory closer to the racket's normal, offering better control and exhibiting superior stability performance. However, these findings diverge from those of (Brannigan & Adali, 1981), who did not measure the ball's rebound angle but relied on simulation results without experimental validation.

Within the tested range, the study indicates that as the racket string tension increases, the contact time between the tennis ball and the strings shows a decreasing trend (Fig. 6), with the 35-pound tension exhibiting longer contact times compared to other tensions, and the 70-pound tension having the shortest contact time. This is consistent with the research of scholars (Baktiar *et al.*, 2020) and colleagues. It may be because when the tennis ball impacts lower tension strings, the deformation of the strings is greater, leading to an extended contact time. Longer contact times result in reduced impact vibrations transmitted to the racket and the

player's hand (Kawazoe *et al.*, 2012), and increasing the contact time between the tennis ball and the strings can enhance a player's control over the ball. Therefore, to improve hitting comfort and control effectiveness, lower tension strings are a good choice (Cross, 2000). Additionally, within the tested range, the study also found that as the incidence angle increases, the contact time between the tennis ball and the strings increases slightly. The reason may be that with an increase in the incidence angle, the tennis ball slides or rolls a longer distance on the contact surface (Goodwill & Haake, 2004). With an increase in incidence velocity, the contact time shows a decreasing trend. The possible reason is that the greater deformation of the strings results in shorter contact times (Baktiar *et al.*, 2020).

In this study, three impact velocities (20 m/s, 25 m/s, 30 m/s) were used, whereas prior studies mostly employed a single one (Baker & Wilson, 1978; Baszczyński *et al.*, 2016). The eight-string tensions selected here differ significantly from the 1–3 used in previous research (Bower & Cross, 2005; Kawazoe *et al.*, 2012; Zhao *et al.*, 2025). Also, the five impact angles were chosen to contrast with the single or fewer angles used before (Bower & Sinclair, 1999; Cross, 1999). This approach helps comprehensively understand the overall trends in the relationship between string tension and coefficients of restitution, angle ratios, and contact times. It also highlights the substantial influence of string tension on tennis racket performance. Lower tensions (e.g., 40 lbs) yield higher coefficients of restitution, smaller rebound angles, and longer contact times, boosting ball speed and control precision (Allen *et al.*, 2016; Chadefaux *et al.*, 2016; Zhao *et al.*, 2025). Selecting optimal tension requires a comprehensive consideration of speed, control, comfort, and performance (Baszczyński *et al.*, 2016). Beginners and intermediate players are advised to choose medium tensions (40 lbs–50 lbs). Advanced players can select based on personal preference and technique, with lower tensions for power-oriented players and higher tensions for control-focused ones.

This study established three levels of incident speed. Within the same level, each launch speed was nearly identical but not exactly equal. The main factors affecting the launch speed were the stretching time and length of the launcher's elastic band, as well as the ball's position inside the launcher. These factors caused slight differences between the set speed and the actual launch speed. During data analysis, the launch speed obtained from the tennis image analysis was used for calculations, which did not affect the results. The relatively low R-squared value observed in this study may be attributed to the uneven stress distribution at the seams resulting from the internal structural composition of tennis balls during impact events. However, this localized phenomenon does not substantially influence the overall trend of the data.

5. Conclusion

Lower string tensions result in a higher COR and rebound speed. Lower string tensions also lead to smaller reflection angles for the same incidence angle, contributing to more stable ball striking. Additionally, lower string tensions correspond to longer contact times between the tennis ball and the racket strings.

Authors' contributions

All authors contributed equally to this work.

Acknowledgments

This study was supported by the Humanities and Social Sciences Research Planning Fund Project of the Ministry of Education (21YJAZH092) and the Humanities and Social Sciences Program of Liaocheng University (321022124).

References

1. Allen, T., Choppin, S., & Knudson, D. (2016). A review of tennis racket performance parameters. *Sports Engineering*, *19*(1), 1–11. https://doi.org/10.1007/s12283-014-0167-x
2. Baker, J.A.W., & Wilson, B.D. (1978). The effect of tennis racket stiffness and string tension on ball velocity after impact. *Research Quarterly. American Alliance for Health, Physical Education and Recreation*, *49*(3), 255–259. https://doi.org/10.1080/10671315.1978.10615532
3. Baktiar, S.B., Sujae, I.H., Kudo, S., Wan Zakariah, W.R., Ong, A., & Hamill, J. (2020). Establishing a method to determine impact force in tennis – a preliminary study. *Journal of Physics: Conference Series*, *1481*, Article 012001. https://doi.org/10.1088/1742-6596/1481/1/012001
4. Bao, L., Sakurai, M., & Nakazawa, M. (2003). Relation between frictional characteristics and the strings' ability to make a tennis ball spin. *Textile Research Journal*, *73*(7), 570–574. https://doi.org/10.1177/004051750307300702
5. Baszczyński, P., Chevrel-Fraux, C., Ficheux, S., Manin, L., & Triquigneaux, S. (2016). Settings adjustment for string tension and mass of a tennis racket depending on the ball characteristics: Laboratory and field testing. *Procedia Engineering*, *147*, 472–477. https://doi.org/10.1016/j.proeng.2016.06.343
6. Bendtsen, K., Rasmussen, K., Hansen, M.B., Fuglsang, T., & Rasmussen, J. (2015). Determining mechanical parameters for spin in tennis strings. *Medicine and Science in Tennis*, *20*(1), 17–24.
7. Bower, R., & Cross, R. (2003). Player sensitivity to changes in string tension in a tennis racket. *Journal of Science and Medicine in Sport*, *6*(1), 120–131. https://doi.org/10.1016/S1440-2440(03)80015-4
8. Bower, R., & Cross, R. (2005). String tension effects on tennis ball rebound speed and accuracy during playing conditions. *Journal of Sports Sciences*, *23*(7), 765–771. https://doi.org/10.1080/02640410400021914
9. Bower, R., & Cross, R. (2008). Elite tennis player sensitivity to changes in string tension and the effect on resulting ball dynamics. *Sports Engineering*, *11*(1), 31–36. https://doi.org/10.1007/s12283-008-0006-z
10. Bower, R., & Sinclair, P. (1999). Tennis racquet stiffness and string tension effects on rebound velocity and angle for an oblique impact. *Journal of Human Movement Studies*, *37*(6), 271–286. http://hdl.handle.net/10453/17085
11. Brannigan, M., & Adali, S. (1981). Mathematical modelling and simulation of a tennis racket. *Medicine and Science in Sports and Exercise*, *13*(1), 44–53. https://europepmc.org/article/med/7219135
12. Chadefaux, D., Rao, G., Le Carrou, J.-L., Berton, E., & Vigouroux, L. (2016). The effects of player grip on the dynamic behaviour of a tennis racket. *Journal of Sports Sciences*, *35*(12), 1155–1164. https://doi.org/10.1080/02640414.2016.1213411
13. Cross, R. (1999). Dynamic properties of tennis balls. *Sports Engineering*, *2*(1), 23–33. https://doi.org/10.1046/j.1460-2687.1999.00019.x
14. Cross, R. (2000). Flexible beam analysis of the effects of string tension and frame stiffness on racket performance. *Sports Engineering*, *3*(2), 111–122. https://doi.org/10.1046/j.1460-2687.2000.00046.x
15. Cross, R. (2003). Measurements of the horizontal and vertical speeds of tennis courts. *Sports Engineering*, *6*(2), 95–111. https://doi.org/10.1007/BF02903531
16. Cross, R., & Bower, R. (2001). Measurements of string tension in a tennis racket. *Sports Engineering*, *4*(3), 165–175.
17. Cross, R., Lindsey, C., & Andruczyk, D. (2000). Laboratory testing of tennis strings. *Sports Engineering*, *3*(4), 219–230. https://doi.org/10.1046/J.1460-2687.2000.00064.X
18. Ghaednia, H., Cermik, O., & Marghitu, D.B. (2015). Experimental and theoretical study of the oblique impact of a tennis ball with a racket. *Proceedings of the Institution of Mechanical Engineers, Part P: Journal of Sports Engineering and Technology*, *229*(3), 149–158. https://doi.org/10.1177/1754337114567490

19. Goodwill, S.R., & Haake, S.J. (2004). Ball spin generation for oblique impacts with a tennis racket. *Experimental Mechanics*, *44*(2), 195–206. https://doi.org/10.1007/BF02428179

20. Haake, S.J., Carré, M.J., & Goodwill, S.R. (2003). The dynamic impact characteristics of tennis balls with tennis rackets. *Journal of Sports Sciences*, *21*(10), 839–850. https://doi.org/10.1080/0264041031000140329

21. Hatch, G.F., Pink, M.M., Mohr, K.J., Sethi, P.M., & Jobe, F.W. (2006). The effect of tennis racket grip size on forearm muscle firing patterns. *The American Journal of Sports Medicine*, *34*(12), 1977–1983. https://doi.org/10.1177/0363546506290185

22. Hennig, E.M. (2007). Influence of racket properties on injuries and performance in tennis. *Exercise and Sport Sciences Reviews*, *35*(2), 62–66. https://doi.org/10.1249/jes.0b013e31803ec43e

23. Kawazoe, Y., Okimoto, K., & Okimoto, K. (2012). Mechanism of tennis racket spin performance. *Journal of System Design and Dynamics*, *6*(2), 200–212. https://doi.org/10.1299/jsdd.6.200

24. Kotze, J., Mitchell, S.R., & Rothberg, S. (2000). The role of the racket in high-speed tennis serves. *Sports Engineering*, *3*(2), 67–84. https://doi.org/10.1046/j.1460-2687.2000.00050.x

25. Miller, S. (2005). Performance measurement of tennis equipment. *Journal of Mechanics in Medicine and Biology*, *5*(2), 217–229. https://doi.org/10.1142/S0219519405001424

26. Zhao, G., Li, C., & Liu, Y. (2025). Effects of different tennis racket string tension on forehand stroke effect and racket dynamic impact. *PLoS ONE*, *20*(1), Article e0317442. https://doi.org/10.1371/journal.pone.0317442

Manuscript received December 21, 2024; accepted for publication April 29, 2025; published online October 14, 2025.

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 865–876, Warsaw 2025
https://doi.org/10.15632/jtam-pl/205901

INFLUENCE OF CONICAL STRUCTURE ON SEALING SPECIFIC PRESSURE UNDER STATIC LOADING

Chun LIU*, Luo Jin LI, Shu Wen HU

Laboratory of Science and Technology on Cryogenic Liquid Propulsion of CASC, Beijing Aerospace Propulsion Institute, Beijing, China

*corresponding author, *lchun1123@163.com*

Similarity analysis and numerical simulations are performed to investigate the effects of axial force, material physical properties, and geometric shape of the conical structure on the sealing specific pressure. The results indicate that under a certain axial force, the conical structure can achieve a high sealing specific pressure. However, the sealing specific pressure decreases with the increase in the sealing surface diameter, sealing surface width, cone angle, and friction coefficient. In terms of material physical properties, the sealing specific pressure increases with the increase in Young's modulus of the upper cone, while other performance parameters have little effect on the sealing specific pressure. In addition, by using similarity analysis, a semi-empirical analytical expression model is proposed to represent the dependence of sealing specific pressure on the axial force, friction coefficient, material physical properties, and geometric properties of the conical structure.

Keywords: conical sealing structure; sealing specific pressure; similarity analysis; numerical simulation.

1. Introduction

Valves, as important components of process fluid systems, are widely used in industries, such as petroleum, chemical, aerospace, etc. Their sealing performance directly affects the effectiveness of the entire pipeline control system, which has been the focus of research (Deng *et al.*, 2023; Chen *et al.*, 2024). There are various types of valve seals, mainly including plane seals, spherical seals, curved seals, and conical seals. And with the development of numerical calculations, scholars have studied the influence of different factors on the sealing pressure of valves (Song & Zheng, 2013; Wang *et al.*, 2024a; Kwak *et al.*, 2019; Peng *et al.*, 2021; Abbasov *et al.*, 2021; Zhao *et al.*, 2022; Yuvaraj & Arunkumar, 2025).

Li *et al.* (2023) conducted the sensitivity analysis of sealing structure parameters and determined the optimum size of the valve. The results show that the sealing performance of the valve was significantly improved after optimization. Yang *et al.* (2020) analyzed the influence of parameters such as the maximum interference fit and taper of the sealing ring contact surface on the sealing contact stress, and found that increasing the interference fit and taper is beneficial for improving the sealing performance. Wu *et al.* (2010) proposed a deep high-pressure conical valve based on a polyether ether ketone seat sealing structure and studied the sealing performance of the new valve. The results show that the new valve has better sealing performance. Jayanath *et al.* (2016) conducted a finite element analysis of the contact stress on nitrile rubber seals for valves, and their experimental results were found to be consistent with the finite element analysis results. Lin *et al.* (2022) analyzed the influence of sealing pair structure dimensions and medium pressure on sealing pressure. Wang *et al.* (2024b) propose a design scheme for an outer conical sealing structure to address the shortcomings of the conical sealing structure. The ad-

vantage of its sealing principle is that it not only ensures the sealing effect, but also reduces processing difficulty. Wu *et al.* (1992) analyzed the force state of the conical sealing pair and derived the relationship between the axial force and cone angle of the sealing pressure. However, this relationship ignores the fact that different conical sealing pairs will undergo different deformations under different axial forces, resulting in changes in the stress state. In addition, some researchers (Gorash *et al.*, 2016; Romanik *et al.*, 2019; Kwak *et al.*, 2019; Li *et al.*, 2024) have also conducted research on the design and performance of valve sealing structures, and obtained the sealing contact pressure under different working conditions.

In summary, researchers have made some progress in the design and performance of valve sealing structures. However, further research is needed on the sealing-specific pressure model of conical structures. On the one hand, although the theoretical formula for sealing specific pressure was established in early stages, the theoretical assumptions are too simplistic due to the deformation of the conical structure under axial force, resulting in significant deviations in calculations. On the other hand, previous simulation studies have only considered the influence of a small number of factors on the sealing specific pressure, making it difficult to obtain the functional relationship between the sealing specific pressure and various factors. Therefore, the purpose of this study is to investigate the quantitative dependence of sealing specific pressure on the axial force, friction coefficient, material physical properties, and geometric properties of the conical structure, so as to further guide the design of valve seal structure.

2. Similarity analysis of sealing specific pressure for conical structure

The sealing problem for a conical structure under axial force is illustrated in Fig. 1. When the upper cone is compressed against the lower cone by an axial force, the two cones undergo certain deformation and form a sealing surface with a certain sealing specific pressure, thereby achieving the sealing effect. In the following analysis, the sealing specific pressure for a conical structure under such axial force is derived using similarity analysis.

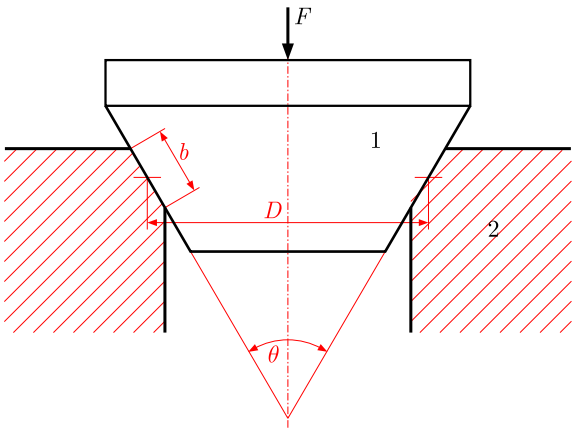

Fig. 1. Schematic diagram of conical structure: 1 – upper cone; 2 – lower cone.

Obviously, the specific sealing pressure for a conical structure depends on the axial force, the physical properties of material, and the geometric properties of the conical structure. Therefore, the sealing specific pressure p can be written as

$$p = f(F, D, b, \theta, \mu, E_1, E_2, v_1, v_2, \rho_1, \rho_2, \sigma_1, \sigma_2), \tag{2.1}$$

where F is the axial force, D is the sealing surface diameter, b is the sealing surface width, θ is the cone angle, μ is the friction coefficient, E_1 and v_1 are Young's modulus and Poisson's ratio of the lower cone, respectively. E_2 and v_2 are Young's modulus and Poisson's ratio of the upper

cone, respectively, ρ_1 and ρ_2 are the densities of the upper cone and lower cone, respectively, σ_1 and σ_2 are the yield strengths of the upper cone and lower cone, respectively.

The sealing problem for a conical structure under axial force is a static problem, so the inertia effect is not considered. The gravity of the upper cone is much smaller than the axial force, so the influence of cone gravity on the sealing specific pressure is negligible. In other words, it can be assumed that p is independent of ρ_1 and ρ_2. Additionally, to ensure the strength reliability of the conical sealing structure, the stresses on both cones should be less than the yield strength of the materials, so the yield strength has almost no effect on the change in the sealing specific pressure, which means that p can be regarded as independent of σ_1 and σ_2. Therefore, Eq. (2.1) can be simplified as

$$p = f(F, D, b, \theta, \mu, E_1, E_2, v_1, v_2). \tag{2.2}$$

In the conical sealing structure, Young's modulus of the lower cone is larger than that of the upper cone. When the upper cone is compressed against the lower cone by axial force, the upper cone undergoes more deformation than the lower cone, causing the area of the contact surface to change, which is consistent with the subsequent simulation results. The changes in Young's modulus and Poisson's ratio of the lower cone in a certain range have little effect on the deformation and thus on the area of the contact surface. Therefore, the influence of Young's modulus and Poisson's ratio for the lower cone on the specific sealing pressure can be ignored. Equation (2.2) can be rewritten as

$$p = f(F, D, b, \theta, \mu, E_2, v_2). \tag{2.3}$$

According to Barenblatt (1996), in the MLT (corresponding to mass, length, and time) class of systems of units, the dimension of each quantity involved in Eq. (2.3) can be defined as follows:

$$\begin{aligned} &\dim p = \mathrm{ML}^{-1}\mathrm{T}^{-2}, && \dim F = \mathrm{MLT}^{-2}, \\ &\dim D = \dim b = \mathrm{L}, && \dim \theta = \dim \mu = 1, \\ &\dim E_2 = \mathrm{ML}^{-1}\mathrm{T}^{-2}, && \dim v_2 = 1. \end{aligned} \tag{2.4}$$

It is straightforward to verify that the dimensions of F and D are independent. By considering these two variables as the basic system determining the independent variable, the remaining dependent variables in Eq. (2.4) can be expressed as

$$\begin{aligned} &\dim p = \dim (F/D^2), && \dim b = \dim D, \\ &\dim \mu = \dim \theta = \dim v_2 = 1, \\ &\dim E_2 = \dim (F/D^2). \end{aligned} \tag{2.5}$$

Therefore, by using the Buckingham Pi theorem, Eq. (2.3) can be written in the following dimensionless form:

$$\Pi = f(\Pi_1, \Pi_2, \Pi_3, \Pi_4, \Pi_5), \tag{2.6}$$

where f is an arbitrary function and

$$\begin{aligned} &\Pi = \frac{p}{F/D^2}, \\ &\Pi_1 = \frac{b}{D}, \quad \Pi_2 = \theta, \quad \Pi_3 = \mu, \quad \Pi_4 = \frac{E_2}{F/D^2}, \quad \Pi_5 = v_2. \end{aligned} \tag{2.7}$$

By substituting Eq. (2.7) into Eq. (2.6), we get

$$\frac{p}{F/D^2} = f\left(\frac{b}{D}, \theta, \mu, \frac{E_2}{F/D^2}, v_2\right). \tag{2.8}$$

One can see that there are now five dimensionless arguments in Eq. (2.8), which can be simplified further by using incomplete similarity or the second type of self-similarity (Barenblatt, 1996).

First, the dimensionless parameter Π_1 is analyzed. The sealing surface width used in this study is $b = 0.4$–$2\,\text{mm}$, the sealing surface diameter is $D = 4$–$20\,\text{mm}$, and there is $\Pi_1 = 0.02$–0.5. Moreover, when other parameters remain constant, the larger the sealing surface width b, the larger the contact area between the two conical surfaces, and the smaller the sealing specific pressure p. According to Barenblatt (1996), in traditional "physical level" discussions, the parameter Π_1 should be considered essential. This indicates that Π may have incomplete self-similarity or similarity of the second type in the dimensionless parameter Π_1. In other words, assuming that a function f_1 has an arbitrary power law-type asymptotic expression, Eq. (1.6) can be written in the following simplified form:

$$\Pi = \Pi_1^{\alpha} f_1(\Pi_2, \Pi_3, \Pi_4, \Pi_5), \tag{2.9}$$

where α is an undetermined constant exponent.

Secondly, regarding the dimensionless parameter Π_2 and Π_3, we have $\Pi_2 = 50°$–$90°$ and $\Pi_3 = 0.3$–0.38 in this study. According to the force analysis of the conical sealing structure, it can be concluded that for Π_2, when other parameters remain constant, as the cone angle increases, the normal stress on the contact surface decreases and the sealing specific pressure decreases, which is consistent with the subsequent simulation results; for Π_3, when other parameters remain constant, as the cone angle increases, the normal stress on the contact surface decreases and the sealing specific pressure decreases, which is consistent with the subsequent simulation results. As in the analysis for Π_1, Π has incomplete self-similarity or similarity of the second type in the dimensionless parameters Π_2 and Π_3. Therefore, Eq. (2.9) can be written in the following simpler form:

$$\Pi = \Pi_1^{\alpha} \Pi_2^{\beta} \Pi_3^{\gamma} f_2(\Pi_4, \Pi_5), \tag{2.10}$$

where f_2 is an arbitrary function, and α, β, γ are three undetermined constant exponents.

Finally, regarding the dimensionless parameter Π_4, with $F = 200\,\text{N}$, $E_2 = 2.67\,\text{GPa}$, and $D = 8\,\text{mm}$, we have $\Pi_4 = 854$, which is much greater than 10. However, when Young's modulus E_2 decreases within a certain range, the increase in contact area leads to an increase in the sealing specific pressure p, which will not approach a constant. According to Barenblatt (1996), this indicates that Π may have incomplete self-similarity or similarity of the second type in the dimensionless parameter Π_4. Then Eq. (2.10) can be expressed as

$$\Pi = \Pi_1^{\alpha} \Pi_2^{\beta} \Pi_3^{\gamma} \Pi_4^{\eta} f_3(\Pi_5), \tag{2.11}$$

where f_3 is an arbitrary function, and α, β, γ, η are four undetermined constant exponents. Substituting Eq. (2.7) into Eq. (2.11) yields

$$\frac{p}{F/D^2} = \left(\frac{b}{D}\right)^{\alpha} \theta^{\beta} \mu^{\gamma} \left(\frac{E_2}{F/D^2}\right)^{\eta} f_3(v_2). \tag{2.12}$$

The function $f_3(v_2)$ and the four undetermined constant exponents α, β, γ ,η will be determined based on the following numerical results.

3. Numerical simulation

3.1. Simulation model

The sealing specific pressure of the conical structure under axial force was investigated using ANSYS-WORKBENCH. The conical structure is an axisymmetric structure, so the model is simplified to a 1/2 model. The upper cone material is aluminum alloy 2A14. The lower cone material is 20Cr13, and the physical properties of the material are shown in Table 1. Additionally, grid partitioning is a crucial step in simulation analysis, where grid quality has a significant impact on the accuracy and precision of the simulation analysis. In order to ensure sufficient accuracy of the calculation results, the mesh of the contact area was encrypted with a mesh size of 0.02 mm, as shown in Fig. 2a. The contact surface is set to frictional contact. The bottom boundary of the lower cone is set as a fixed support to prevent the lower cone from moving. The displacement constraint of the upper cone surface in the X- and Z-directions is 0 mm, and only axial movement is allowed. Axial force is applied to the upper cone surface to ensure sealing, as shown in Fig. 2b.

Table 1. Physical properties of material.

Material	ρ [kg/m^3]	E [GPa]	v	σ_s [MPa]
2A14	2800	72	0.33	380
20Cr13	7750	200	0.3	540

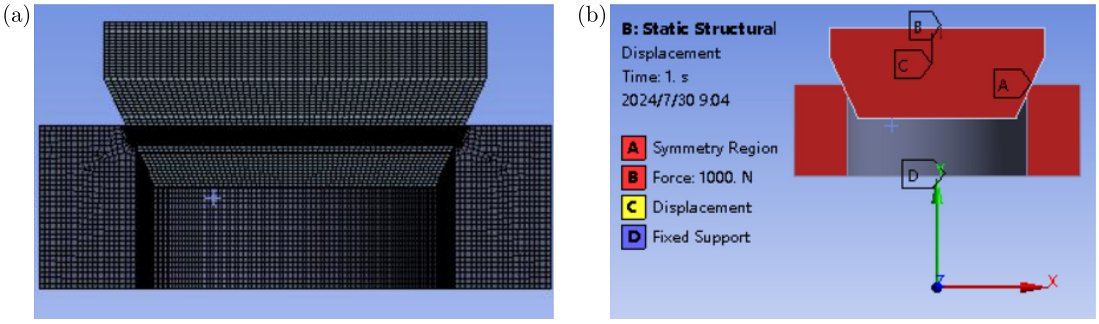

Fig. 2. Simulation model of cone seal structure: (a) grid distribution; (b) boundary conditions.

3.2. Numerical results and discussion

3.2.1. Dependence of sealing specific pressure on axial force

To study the effects of axial force on sealing specific pressure, for an axial force range of 500 N to 2000 N, we numerically simulated a conical structure with a diameter of $D = 12$ mm, width of $b = 0.8$ mm, and angle of $\theta = 50°$. In all of these numerical simulations, the friction coefficient was $\mu = 0.2$.

Figure 3 shows the equivalent stress distribution of the conical structure under an axial force $F = 2000$ N. In Fig. 3, it can be seen that the maximum equivalent stresses of the upper and

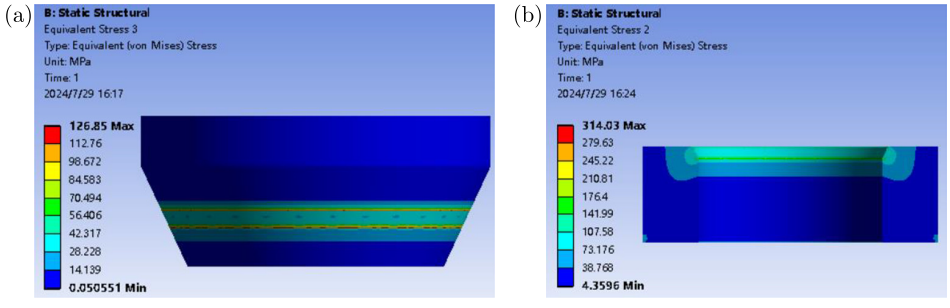

Fig. 3. Equivalent stress distribution of conical sealing structure: (a) upper cone; (b) lower cone.

lower cones are 126.8 MPa and 314 MPa, respectively, both of which are lower than the yield strength of the material. In addition, the equivalent stress in the middle part of the contact surface is relatively small, while the equivalent stress in the upper and lower boundary parts is relatively large. Figure 4 shows the equivalent strain distribution of the conical structure under an axial force $F = 2000$ N. In Fig. 4, it can be seen that the maximum strains of the upper and lower cones are 0.0018 mm/mm and 0.0016 mm/mm, respectively, and the deformation of the upper conical surface is relatively large. Figure 5 shows the sealing-specific pressure distribution on the sealing contact surface under an axial force of $F = 2000$ N. In Fig. 5, it can be seen that the sealing specific pressure of the conical structure is relatively small in the middle of the sealing band, with the highest values at the upper and lower boundary positions, about 298 MPa, and the positions of the maximum stress and maximum deformation correspond to the positions of the maximum sealing specific pressure.

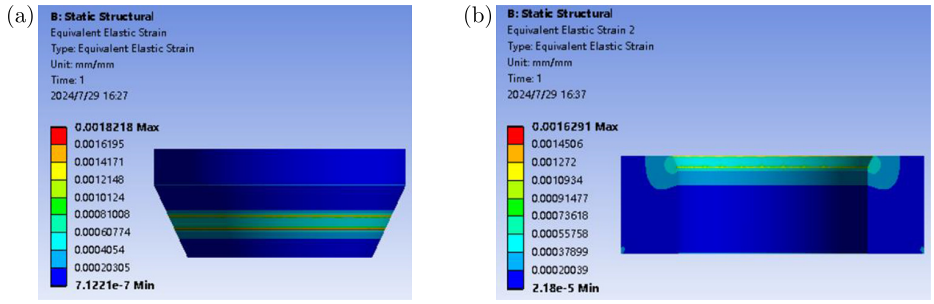

Fig. 4. Strain distribution of conical sealing structure: (a) upper cone; (b) lower cone.

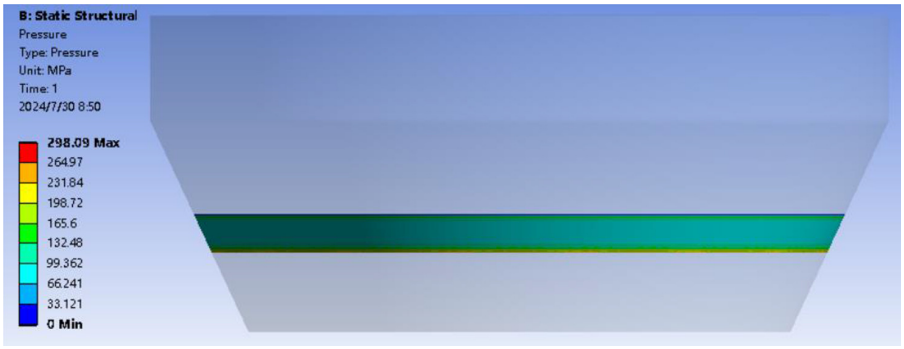

Fig. 5. Sealing specific pressure on the contact surface.

To obtain a general rule for the dependence of the sealing specific pressure on the axial force, additional numerical simulations with various values of F were conducted. The numerical results are presented in Table 2. One can observe that with an increase in the axial force, the stress and sealing specific pressure of the conical structure also increase. Under the action of the axial force $F = 2500$ N, the stress of the conical structure with $D = 12$ mm, $b = 0.8$ mm, and $\theta = 50°$ is less than the yield strength of the material, indicating high reliability.

Table 2. Sealing specific pressure of conical sealing structure under different axial forces.

F [N]	Upper cone stress [MPa]	Lower cone stress [MPa]	p [MPa]
500	36.7	99.1	87.7
1000	67.2	184.4	162.2
1500	98.1	270.1	236.6
2000	126.8	314.0	298.1
2500	178.3	367.4	355.4

3.2.2. Dependence of sealing specific pressure on sealing surface diameter

To study the effects of the sealing surface diameter on sealing specific pressure, numerical simulations for conical sealing structures with various sealing surface diameters ($D = 4$ mm, 8 mm, 12 mm, 16 mm, and 20 mm) were conducted with the width of $b = 0.8$ mm, and angle of $\theta = 70°$. In all of these numerical simulations, the axial force was $F = 2000$ N and the friction coefficient was $\mu = 0.2$.

The sealing specific pressure for different sealing diameters is shown in Table 3. According to Table 3, as the sealing surface diameter increases, the stress and sealing specific pressure of the conical structure gradually decrease. The sealing specific pressure of the conical structure with a sealing surface diameter of $D = 4$ mm is the highest, at 640 MPa, which is 3.1 times the sealing specific pressure of the conical structure with a sealing surface diameter of $D = 20$ mm. This is because, under the same axial force, a smaller sealing surface diameter results in a smaller contact area, leading to a higher sealing specific pressure.

Table 3. Sealing specific pressure of conical structures with different sealing surface diameters.

D [mm]	Upper cone stress [MPa]	Lower cone stress [MPa]	p [MPa]
4	294.4	627.0	660.0
8	145.8	325.3	371.0
12	99.2	227.5	264.9
16	80.0	183.8	221.9
20	69.6	163.5	205.6

3.2.3. Dependence of sealing specific pressure on sealing surface width

To study the effects of the sealing surface width on sealing specific pressure, numerical simulations for conical structures with various sealing surface widths ($b = 0.4$ mm, 0.8 mm, 1.2 mm, 1.6 mm, and 2 mm) were conducted with a diameter of $D = 0.8$ mm, and an angle of $\theta = 70°$. In all of these numerical simulations, the axial force was $F = 2000$ N and the friction coefficient was $\mu = 0.2$.

The numerical results of sealing specific pressure for conical structures with different sealing surface widths are shown in Table 4. In Table 4, one can see that as the sealing surface width increases, the stress and sealing specific pressure of the conical structure gradually decrease. This is because under the same axial force, a smaller sealing surface width will result in a smaller contact area, leading to a higher sealing specific pressure.

Table 4. Sealing specific pressure of conical structures with different sealing surface widths.

b [mm]	Upper cone stress [MPa]	Lower cone stress [MPa]	p [MPa]
0.4	138.8	326.4	348.7
0.8	99.2	227.5	264.9
1.2	84.2	214.5	223.6
1.6	75.6	173.6	207.4
2	68.9	160.1	195.6

3.2.4. Dependence of sealing specific pressure on cone angle

To study the effects of the cone angle on sealing specific pressure, numerical simulations for conical structures with various cone angles ($\theta = 50°$, $60°$, $70°$, $80°$, and $90°$) were conducted with

the diameter of $D = 0.8\,\text{mm}$, and width of $b = 12\,\text{mm}$. In all of these numerical simulations, the axial force was $F = 2000\,\text{N}$ and the friction coefficient was $\mu = 0.2$.

The numerical results of sealing specific pressure for conical structures with different cone angles are shown in Table 5. According to Table 5, the stress and sealing specific pressure of the conical structure gradually decrease with the increase in the cone angle. This is because under the same axial force, the increase in the cone angle reduces the normal stress on the contact surface, resulting in a decrease in sealing specific pressure, which is consistent with the theoretical analysis results.

Table 5. Sealing specific pressure of conical structures with different cone angles.

θ [°]	Upper cone stress [MPa]	Lower cone stress [MPa]	p [MPa]
50	126.8	314.1	298.1
60	111.2	284.1	286.5
70	99.2	227.5	264.9
80	89.2	218.2	246.2
90	81.3	209.1	235.6

3.2.5. Dependence of sealing specific pressure on friction coefficient

To study the effects of the friction coefficient on sealing specific pressure, numerical simulations for conical sealing structures with various friction coefficients ($\mu = 0.1$, 0.2, 0.3, 0.4, and 0.5) were conducted with a diameter of $D = 0.8\,\text{mm}$, a cone angle of $\theta = 50°$, and a width of $b = 8\,\text{mm}$. In all of these numerical simulations, the axial force was $F = 2000\,\text{N}$.

The numerical results of sealing specific pressure for conical structures with different friction coefficients are shown in Table 6. According to Table 6, the stress and sealing specific pressure of the conical structure gradually decreases with the increase in the friction coefficient. This is because under the same axial force, the increase in the friction coefficient reduces the normal stress on the contact surface, resulting in a decrease in sealing specific pressure, which is consistent with the theoretical analysis results.

Table 6. Sealing specific pressure of conical structures with different friction coefficients.

μ	Upper cone stress [MPa]	Lower cone stress [MPa]	p [MPa]
0.1	157.2	361.3	336.2
0.2	126.8	314.0	298.0
0.3	128.3	328.9	270.7
0.4	130.5	308.6	240.1
0.5	135.7	291.7	215.4

3.2.6. Dependence of sealing specific pressure on Young's modulus and Poisson's ratio

To study the effects of Young's modulus and Poisson's ratio on sealing specific pressure, numerical simulations for conical sealing structures were conducted with the diameter of $D = 0.8\,\text{mm}$, cone angle of $\theta = 50°$, and width of $b = 8\,\text{mm}$. In all of these numerical simulations, the axial force was $F = 2000\,\text{N}$ and the friction coefficient was $\mu = 0.2$.

The influence of Young's modulus ($E_2 = 100\,\text{GPa}$, 120 GPa, 150 GPa, 180 GPa, and 200 GPa) of the upper cone on the sealing specific pressure was first analyzed. The numerical results of sealing specific pressure for different Young's modulus of the upper cone are shown in Table 7. According to Table 7, the sealing specific pressure of the conical structure gradually increases

Table 7. Sealing specific pressure of conical structures with different Young's modulus.

E_2 [GPa]	Upper cone stress [MPa]	Lower cone stress [MPa]	p [MPa]
100	135.8	370.8	325.6
120	139.6	377.8	333.6
150	144.9	389.2	347.5
180	150.0	399.4	358.3
200	153.2	405.7	364.8

with the increase in Young's modulus of the upper cone. This is because under the same axial force, the increase in Young's modulus of the upper cone causes less deformation, resulting in a smaller contact area and an increase in sealing specific pressure. In addition, the influence of Poisson's ratio ($v_2 = 0.3$, 0.32, 0.34, 0.36, and 0.38) of the upper cone on the sealing specific pressure was analyzed. The numerical results of sealing specific pressure for different Poisson's ratios are shown in Table 8. According to Table 8, with the increase in Poisson's ratio of the upper cone, the change in sealing specific pressure is relatively small, at approximately 4 MPa, which can be almost ignored.

Table 8. Sealing specific pressure of conical structures with different Poisson's ratio.

v_2	Upper cone stress [MPa]	Lower cone stress [MPa]	p [MPa]
0.3	156.9	407.1	366.2
0.32	154.4	406.2	365.3
0.34	152.0	405.2	364.4
0.36	149.8	404.1	363.5
0.38	147.7	403.1	362.6

4. Semi-empirical analytical expression for sealing specific pressure

To determine the model constants α, β, γ, η, and the function $f_3(v_2)$ in Eq. (2.12), the numerical results derived above are used. Numerical fitting of these numerical results will be conducted using the Levenberg–Marquardt optimization algorithm.

Firstly, in Table 8, one can see that within a certain range, the change in Poisson's ratio of the upper cone has a relatively small impact on the sealing specific pressure and can be ignored. Therefore, when other parameters (axial force, sealing surface diameter, sealing surface width, cone angle, friction coefficient, Young's modulus of upper cone) remain unchanged, $f_3(v_2)$ can be written in the following form:

$$f_3(v_2) = c, \tag{4.1}$$

where c is a constant. Therefore, substituting Eq. (4.1) into Eq. (2.12) yields

$$\frac{p}{F/D^2} = c\left(\frac{b}{D}\right)^{\alpha}\theta^{\beta}\mu^{\gamma}\left(\frac{E_2}{F/D^2}\right)^{\eta}. \tag{4.2}$$

The parameters in Eq. (4.2) were determined by fitting the above simulation results. The fitting results and model parameters are shown in Fig. 6 and Table 9, respectively. Therefore, the sealing specific pressure of the conical structure can ultimately be expressed as

$$p = \frac{0.72F^{0.79}E_2^{0.21}}{D^{1.12}b^{0.46}\theta^{0.44}\mu^{0.25}}. \tag{4.3}$$

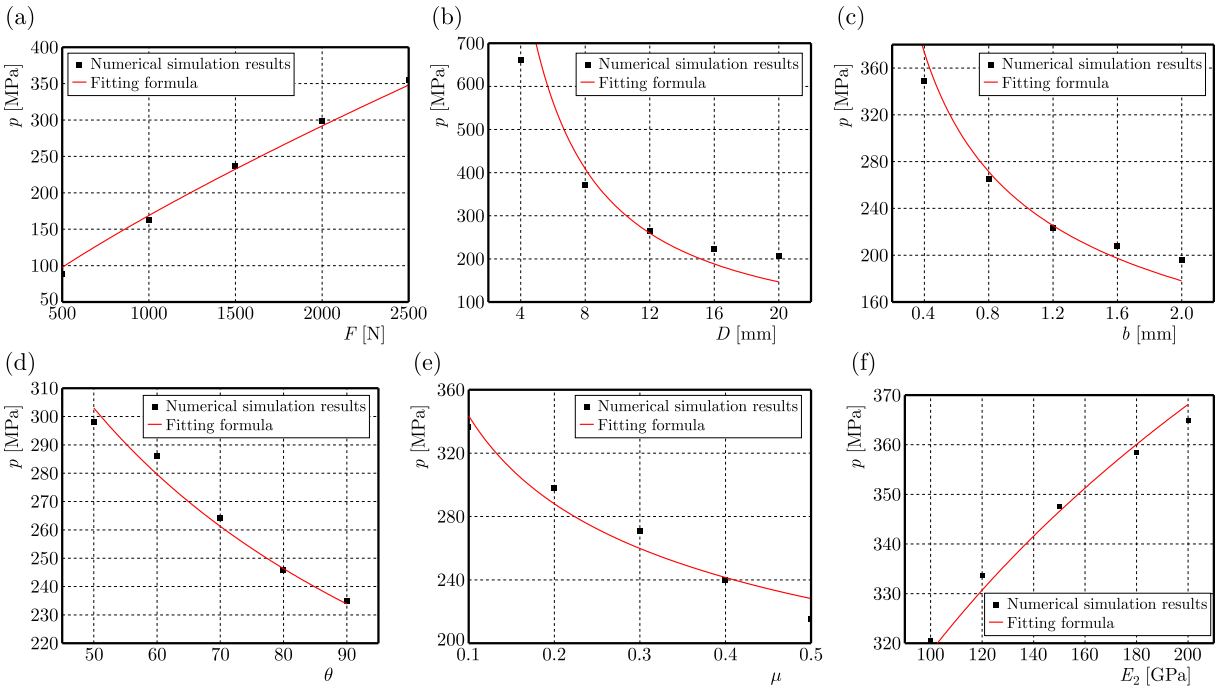

Fig. 6. Sealing specific pressure versus (a) axial force, (b) sealing surface diameter, (c) sealing surface width, (d) cone angle, (e) friction coefficient, and (f) Young's modulus.

Table 9. Parameters of the model.

c	α	β	γ	η
0.72	−0.46	−0.44	−0.25	0.21

5. Conclusions

In this study, similarity analysis and numerical simulations were conducted on the relationship between the sealing specific pressure and various factors for the conical structure under axial loading, specifically involving axial force, cone angle, sealing surface width, sealing surface diameter, and physical properties of materials. The main conclusions are as follows:

1) based on the similarity theory, a semi-empirical analytical expression for the sealing specific pressure of the conical structure was obtained, which can well predict the dependence of the sealing specific pressure on axial force, material physical properties, and the geometric shape of the conical structure (cone angle, sealing surface width, sealing surface diameter);
2) in addition to the influence of the geometric shape of the conical structure on the sealing specific pressure, the friction coefficient also has a significant impact on the sealing specific pressure, which requires us to ensure that the conical surface has appropriate roughness when being processed;
3) in the metal conical structure, Young's modulus of the upper cone has a significant impact on the sealing specific pressure, which has guiding significance for the selection of conical structure materials.

References

1. Abbasov, E.M., Rustamova, K.O., & Darishova, A.O. (2021). Influence of viscous-elastic properties of a cylindrical sealing element on its sealing ability. *Journal of Theoretical and Applied Mechanics*, *59*(3), 481–492. https://doi.org/10.15632/jtam-pl/140228

2. Barenblatt, G.I. (1996). *Scaling, self-similarity, and intermediate asymptotics*. Cambridge University Press. https://lab.semi.ac.cn/library/upload/files/2021/8/30133839528.pdf

3. Chen, Y.Q., Yu, H., Fu, S.W., & Li, A. (2024). Study on sealing performance of main gate valve of marine nuclear power plant based on fluid-solid-thermal coupling method. *Frontiers in Energy Research*, *12*, Article 1375806. https://doi.org/10.3389/fenrg.2024.1375806

4. Deng, C., Nie, J., Luo, L., Wang, B., Chen, G., Wu, S., Shi, Y., Shi, J., Gao, W., Liu, S., & Zhang, W. (2023). Experimental analysis of the effect of sealing surface pitting defects on valve sealing performance. In H. Dong, & H. Yu (Eds.). *Proceedings: Vol. 12801. Ninth International Conference on Mechanical Engineering, Materials, and Automation Technology (MMEAT 2023)* (Article 128011V). SPIE. https://doi.org/10.1117/12.3007569

5. Gorash, Y., Dempster, W., Nicholls, W.D., Hamilton, R., & Anwar, A.A. (2016). Study of mechanical aspects of leak tightness in a pressure relief valve using advanced FE-analysis. *Journal of Loss Prevention in the Process Industries*, *43*, 61–74. https://doi.org/10.1016/j.jlp.2016.04.009

6. Jayanath, S., Achuthan, A., Mashue, A., & Huang, M. (2016). A subscale experimental test method to characterize extrusion-based elastomer seals. *Journal of Tribology*, *138*(3), Article 032201. https://doi.org/10.1115/1.4032175

7. Kwak, H.-S., Seong, H., & Kim, C. (2019). Design of laminated seal in cryogenic triple-offset butterfly valve used in LNG marine engine. *International Journal of Precision Engineering and Manufacturing*, *20*(2), 243–253. https://doi.org/10.1007/s12541-019-00056-6

8. Li, S., Kang, W., Liu, T., Yang, L., & Wang, Y. (2023). Sealing structure optimization of LNG ultra-low temperature and high-pressure axial flow check valve. *Journal of Applied Science and Engineering*, *26*(7), 1013–1024. https://doi.org/10.6180/jase.202307_26(7).0012

9. Li, S.X., Zheng, M.X., Wang, Y.X., Yang, L.X., & Ma, T.Q. (2024). Multi-objective optimization design of double resilient groove metal seat for ball valve in liquid hydrogen receiving stations. *Journal of the Brazilian Society of Mechanical Sciences and Engineering*, *46*(1), Article 34. https://doi.org/10.1007/s40430-023-04602-2

10. Lin, Z.-H., Yu, L.-J., Hua, T.-F., Jin, Z.-J., & Qian, J.-Y. (2022). Seal contact performance analysis of soft seals on high-pressure hydrogen charge valves. *Journal of Zhejiang University-SCIENCE A (Applied Physics & Engineering)*, *23*(4), 247–256. https://doi.org/10.1631/jzus.A2100395

11. Peng, C., Fischer, F.J., Schmitz, K., & Murrenhoff, H. (2021). Comparative analysis of leakage calculations for metallic seals of ball-seat valves using the multi-asperity model and the magnification-based model. *Tribology International*, *163*, Article 107130. https://doi.org/10.1016/j.triboint.2021.107130

12. Romanik, G., Jaszak, P., & Rogula, J. (2019). Cooperation of the PTFE sealing ring with the steel ball of the valve subjected to durability test. *Open Engineering*, *9*(1), 321–328. https://doi.org/10.1515/eng-2019-0028

13. Song, R., & Zheng, F. (2013). Analysis of ball valve sealing pressure ratio based on UG Nastran. *Advanced Materials Research*, *703*, 204–207. https://doi.org/10.4028/www.scientific.net/AMR.703.204

14. Wang, C.L., Xu, D.T., Huang, K.X., Liu, Y., & Yang, L. (2024a). Multi-objective optimization of a triple-eccentric butterfly valve considering structural safety and sealing performance. *ISA Transactions*, *155*, 295–308. https://doi.org/10.1016/j.isatra.2024.10.009

15. Wang, X., Cao, J., Yang, B., & Wu, X.-J. (2024b). Design and application of a new sealing structure (in Chinese). *Hydraulics Pneumatics & Seals*, *44*(12), 118–122. https://doi.org/10.3969/j.issn.1008-0813.2024.12.019

16. Wu, H.J., Guo, B.Z., & Ding, Y.R. (1992). Theoretical discussion on the selection of cone angle of valve cone sealing (in Chinese). *Fluid Machinery*, *09*, 22–24.

17. Wu, S.-J., Yang, C.-J., Chen, Y., & Xie, Y.-Q. (2010). A study of the sealing performance of a new high-pressure cone valve for deep-sea gas-tight water samplers. *Journal of Pressure Vessel Technology*, *132*(4), Article 041601. https://doi.org/10.1115/1.4001204

18. Yang, Y.W., Zhu, H.W., He, D.S., Du, C.C., Xu, L.B., He, Y., Zheng, Y., & Ye, Z.W. (2020). Contact mechanical behaviors of radial metal seal for the interval control valve in intelligent well: Modeling and theoretical study. *Energy Science & Engineering*, *8*(4), 1337–1352. https://doi.org/10.1002/ese3.597

19. Yuvaraj, K., & Arunkumar, G. (2025). Simulation and validation of the effect of hydrostatic and non-hydrostatic pressure on contact pressure in a resilient seat butterfly valve. *Materials Research Express*, *12*(1), Article 016511. https://doi.org/10.1088/2053-1591/ada877

20. Zhao, B., Zhang, S., Gao, Q., Miao, C., & Wang, T. (2022). Optimization of sealing parameters of double-sealing pipeline repair clamp. *Journal of Theoretical and Applied Mechanics*, *60*(3), 333–346. https://doi.org/10.15632/jtam-pl/150679

Manuscript received April 10, 2025; accepted for publication June 4, 2025; published online October 28, 2025.

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 877–887, Warsaw 2025
https://doi.org/10.15632/jtam-pl/206912

ANALYSIS OF MECHANOCHEMICAL DIFFUSION COUPLING PROCESSES BASED ON TRANSIENT CONTINUUM CHEMO-MECHANICAL COUPLING THEORY

Pengfei YU*, Wenchao SUN, Yaohong SUO, Yihan WU

School of Mechanical Engineering and Automation, Fuzhou University, Fuzhou, China
*corresponding author, *yupengfei0422@fzu.edu.cn*

Chemical diffusion is vital in materials science and energy technology. Current Fick and non-Fick theories overlook the transient nature of diffusion. By referring to biomechanical axioms, we incorporate the transient expansion process and introduce characteristic time. This paper explores chemo-mechanical coupling in a spherical structure via transient continuum theory. The results show characteristic time changes in the diffusion equation from parabolic to hyperbolic, yielding finite diffusion speed and wave-like behavior, offering a basis for optimizing systems like lithium-ion batteries.

Keywords: mechano-diffusion coupling; transient diffusion; characteristic time.

1. Introduction

Chemical diffusion refers to the directional migration of atoms or molecules of a substance due to the presence of a chemical potential gradient. In a system, when the chemical potential varies in different regions, the substance will diffuse from the region with a higher chemical potential to that with a lower chemical potential, aiming to achieve chemical potential equilibrium. As this phenomenon is ubiquitous in almost all materials and fields, the study of chemical diffusion in objects has always attracted extensive interest among researchers (Dai & Xiao, 2021; Hu *et al.*, 2020). The study of this phenomenon has traditionally been based on Fourier's law of heat conduction and Fick's law of diffusion. Both of these laws assume that the propagation speed is infinite. Transient phenomena occurring under high temperatures or extremely large diffusion fluxes, accompanied by mass and heat transfer during chemical reactions, may lead to errors that contradict physical observations. Therefore, it is necessary to discuss the transient effects at finite speeds.

The research on non-Fourier heat conduction has inspired the study of non-Fick diffusion. In light of the analogous nature of heat conduction and mass diffusion, Dong and Jiang (1995) conducted an analysis of the physical process underlying diffusion mass transfer. They introduced the concept of mass propagation speed and formulated the diffusion differential equation for scenarios where the mass propagation speed is finite. This equation addresses the non-Fick diffusion mechanism, which deviates from the traditional Fick diffusion model, and provides a framework for understanding diffusion processes with finite propagation speeds, thereby enhancing the realism of the model in the context of mass transfer phenomena. Liu (2007) employed the least square method and Laplace transform to prognosticate the characteristic time of sodium chloride diffusion in aqueous solutions. This approach was facilitated by leveraging historical experimental data pertaining to the concentration of NaCl diffusion in water. Through computational analysis, Liu successfully obtained theoretical values that were in concordance

with the experimental data, thereby validating the finite propagation speed model for diffusion processes. Kuang (2014) integrated the generalized inertial entropy theory into the framework of continuum thermodynamics, thereby establishing a comprehensive theoretical model. This model elucidates that both heat and diffusion waves propagate at a finite speed, a departure from the infinite propagation speed implied by classical Fourier's law. Suo and Shen (2013a) employed the method of separation of variables to derive the double Fourier series solution for two-dimensional non-Fick diffusion, accommodating arbitrary initial and periodic boundary conditions. This theoretical framework was subsequently validated against both simulation and experimental results, demonstrating a heightened congruence with experimental data when utilizing non-Fick diffusion models. This finding underscores the superiority of non-Fick diffusion models in accurately capturing the complexities of mass transfer phenomena, particularly in scenarios where traditional Fick models fall short.

Diffusion processes are also indeed intricately linked with various physical phenomena, including stress or deformation fields (Yan *et al.*, 2024), temperature fields (Nguyen *et al.*, 2019), chemical reactions (Chen *et al.*, 2023), and electric fields (Yu *et al.*, 2016; Yu & Shen, 2014). This coupling is not merely coincidental but reflects the multifaceted nature of transport phenomena in materials science and engineering. Chu and Lee (1994) have indeed conducted research on the effect of stress on chemical diffusion, emphasizing the intricate relationship between mechanical stress and mass transport phenomena. Their work contributes to the understanding of how chemical stresses can influence diffusion processes, a topic of significant importance in materials science and engineering. Building upon the foundation of irreversible thermodynamics, inertial entropy, and inertial concentration, Hu and Shen (2013) have proposed variational principles for the chemical Gibbs function, Helmholtz function, and internal energy. These principles provide a comprehensive theoretical framework for describing fully coupled thermo-mechanical-chemical problems. Konica and Sain (2020) have indeed developed a thermodynamically consistent continuum model that addresses the high-temperature oxidation of polymers, incorporating the complex coupling between diffusion, chemical reactions, and large deformation of polymers. In the context of mechano-thermo-electro-chemical coupling, Yu and Shen (2014) proposed the variational principle of the fully coupled thermal electrical chemical mechanical problem based on irreversible thermodynamics, and derived the fully coupled governing equations including heat conduction, mass diffusion, electrochemical reaction, and electrostatic potential, which can be used to deal with the coupling problem in solids. Suo and Shen (2013b) also established the non-Fick diffusion and stress coupling equations under one-dimensional conditions, and derived approximate analytical solutions for concentration, stress, and displacement using the Laplace transform and the inverse Laplace transform. Considering the microscopic time and chemo-mechanical coupling effect, Shen (2022) introduced the second-order rate and characteristic time through Taylor expansion to describe the transient process and derived the transient Reynolds transport theorem. Based on the conservation laws, the transient field equations, including mechanical and chemical contributions and microscopic time, were derived, providing a more accurate theoretical framework for studying transient phenomena in complex material systems.

The transient continuum mechano-chemical coupling theory naturally derives finite-speed diffusion equations from conservation laws by introducing characteristic time and second-order rate terms, overcoming the empirical parameter dependence limitation of non-Fick diffusion models. Its core advantages lie in multi-physical field coupling capability and physical self-consistency at microscopic time scales. The objective of our work is to quantitatively investigate the influence of characteristic time on mechano-chemical coupling theory, with systematic comparisons to both classical Fick diffusion and non-Fick diffusion models. The structure of this paper is organized as follows: in Section 2, we introduce the transient continuum chemo-mechanical coupling theory and establish a chemo-diffusion coupling model for lithium ions in a spherical structure. Section 3 discusses the effects of characteristic time and boundary concentration. Finally, we present our conclusions in Section 4.

2. Transient continuum chemo-mechanical coupling theory

By applying the axioms of biomechanics to the transient chemo-mechanical coupling process, and excluding chemical reactions, the mass conservation equation and momentum conservation equation can be derived as follows (Shen, 2022):

$$\frac{\partial c_N}{\partial t} + \frac{t_c}{2}\frac{\partial^2 c_N}{\partial t^2} + \nabla \mathbf{J_N} = 0, \qquad \frac{\partial \rho \mathbf{v}}{\partial t} + \frac{t_c}{2}\frac{\partial^2 \rho \mathbf{v}}{\partial t^2} = \nabla\left(\boldsymbol{\sigma} - \boldsymbol{\sigma}^{\mathbf{V}}\right) + \mathbf{b}, \tag{2.1}$$

where c_N is the particle concentration of particle N, t_c is the characteristic time, $\mathbf{J_N}$ is the particle diffusion flux of particle N, ρ is the system density, $\mathbf{v}$ is the velocity, $\mathbf{b}$ is the body force, $\boldsymbol{\sigma}$ is the Cauchy stress, and $\boldsymbol{\sigma}^{\mathbf{V}}$ is the residual stress caused by convective diffusion or chemical reaction, and its expression is

$$\sigma_{ik}^{V} = \left(\rho v_i + t_c \frac{\partial \rho v_i}{\partial t}\right) v_k. \tag{2.2}$$

If the characteristic time t_c is taken as 0 and the convective diffusion of particles is not considered, the above control equations become the classical mass conservation equation and mechanical momentum conservation equation:

$$\frac{\partial c_N}{\partial t} + \nabla \mathbf{J_N} = 0, \qquad \frac{\partial \rho \mathbf{v}}{\partial t} = \nabla \sigma + \mathbf{b}. \tag{2.3}$$

The spherical structure with radius r_0 is shown in Fig. 1. And in this manuscript, the lithium ions are assumed to diffuse inside the spherical structure.

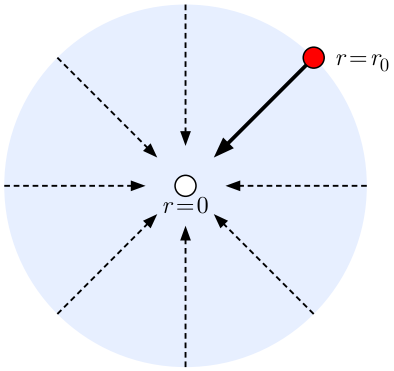

Fig. 1. Schematic diagram of the spherical diffusion model.

Under the influence of concentration, diffusion-induced strain will occur within the sphere. According to Li (1978), the diffusion-induced strain can be expressed as

$$\varepsilon = \Omega(c - c_0), \tag{2.4}$$

where Ω represents the partial molar volume of the sphere, and c_0 represents the initial concentration within the sphere.

Assuming that the deformation in the sphere is relatively small and elastically linear, the material of the spherical model is isotropic, and ions diffuse only radially within the sphere (Suo *et al.*, 2021; 2024). Additionally, the sphere only experiences radial and hoop stresses during ion diffusion, with no shear stress. In a spherical coordinate system, the strain can be expressed in terms of displacement as follows:

$$\varepsilon_r = \frac{\partial u}{\partial r}, \qquad \varepsilon_\theta = \frac{u}{r}, \tag{2.5}$$

where ε_r and ε_θ represent the radial and hoop stresses, respectively, and u is the radial displacement of the sphere.

In spherical coordinates, the constitutive equations for stress and strain under the influence of force-diffusion coupling are given by the following relations (Zhang *et al.*, 2019):

$$\varepsilon_r = \frac{1}{E}\left(\sigma_r - 2\nu\sigma_\theta\right) + \frac{1}{3}\varepsilon, \qquad \varepsilon_\theta = \frac{1}{E}\left[\sigma_\theta - \nu\left(\sigma_\theta + \sigma_r\right)\right] + \frac{1}{3}\varepsilon, \tag{2.6}$$

where σ_r and σ_θ represent the radial and hoop stresses, respectively, while E and ν represent the elastic modulus and Poisson's ratio, respectively.

Assuming that body forces are negligible during the diffusion process, according to Shen (2022), the equilibrium equation is given by:

$$\nabla(\sigma - \sigma^V) = 0. \tag{2.7}$$

Under the influence of stress, the chemical potential of ion diffusion is given by Chu and Lee (1994):

$$\mu = \mu_0 + RT\ln\frac{c}{c_0} - \Omega\sigma_h, \tag{2.8}$$

where $\sigma_h = (\sigma_r + 2\sigma_\theta)/3$ is the hydrostatic pressure.

The velocity v_b of ions during diffusion can be expressed as

$$v_b = -M\nabla\mu = -\frac{D}{c}\frac{\partial c}{\partial r} + \frac{D\Omega}{RT}\frac{\partial\sigma_h}{\partial r}, \tag{2.9}$$

M represents the ionic mobility, ∇ is the gradient operator, and $D = MRT$ is the diffusion coefficient of the material.

Thus, the residual stress can be expressed as

$$\sigma^V = \rho\left(-\frac{D}{c}\frac{\partial c}{\partial r} + \frac{D\Omega}{RT}\frac{\partial\sigma_h}{\partial r}\right)^2. \tag{2.10}$$

Due to the assumption that ions diffuse only in the radial direction, residual stress is generated only in the radial direction within the sphere. Therefore, the equilibrium equation in spherical coordinates is

$$\frac{\partial\left(\sigma_r - \sigma^V\right)}{\partial r} + \frac{2}{r}\left(\sigma_r - \sigma^V - \sigma_\theta\right) = 0. \tag{2.11}$$

Assuming the corresponding boundary conditions are

$$u|_{r=0} = 0, \qquad \sigma_r - \sigma^V\big|_{r=r_0} = 0. \tag{2.12}$$

Under the influence of characteristic time and stress, the diffusion flux J can be expressed as

$$J = \left(c + t_c\frac{\partial c}{\partial t}\right)v_k = -D\left(\frac{\partial c}{\partial r} + \frac{t_c}{c}\frac{\partial c}{\partial t}\frac{\partial c}{\partial r} - \frac{c\Omega}{RT}\frac{\partial\sigma_h}{\partial r} - \frac{t_c\Omega}{RT}\frac{\partial c}{\partial t}\frac{\partial\sigma_h}{\partial r}\right). \tag{2.13}$$

When $t_c = 0$, Eq. (2.13) represents the normal force-diffusion coupling flux under normal conditions.

Ions in the diffusion process obey the law of mass conservation. The mechano-diffusion coupling equation in spherical coordinates is given by

$$\frac{t_c}{2}\frac{\partial^2 c}{\partial t^2} + \frac{\partial c}{\partial t} + \frac{1}{r^2}\nabla\left(r^2 J\right) = 0. \tag{2.14}$$

Based on the symmetry of the sphere, the diffusion flux at the center of the sphere is 0. Assuming that the concentration on the outer surface of the sphere is constant at c_1 and the initial concentration of ions inside the sphere is constant at c_0, the boundary and initial conditions are

$$J|_{r=0} = 0, \qquad c|_{r=r_0} = c_1, \qquad c|_{t=0} = c_0, \qquad \left.\frac{\partial c}{\partial t}\right|_{t=0} = 0. \tag{2.15}$$

To facilitate the simulation calculations, these equations are nondimensionalized by introducing the following dimensionless variables:

$$\begin{aligned} &\overline{r} = \frac{r}{r_0}, && \overline{c} = \frac{c}{c_1}, && \overline{u} = \frac{u}{r_0}, \\ &\overline{\sigma}_h = \frac{\Omega\sigma_h}{RT}, && \overline{f} = \frac{\rho D^2}{E r_0^2}, && \overline{q} = \Omega c_0, \\ &\overline{\sigma}_r = \frac{\Omega\sigma_r}{RT}, && \overline{\sigma}_\theta = \frac{\Omega\sigma_\theta}{RT}, && \overline{t} = \frac{Dt}{r_0^2}, \\ &\overline{\tau} = \frac{Dt_c}{r_0^2}, && \overline{J} = \frac{Jr_0}{Dc_1}. \end{aligned} \tag{2.16}$$

Substituting these dimensionless variables into Eqs. (2.11) and (2.14), the governing equations becomes

$$\begin{aligned} \frac{\partial^2\overline{u}}{\partial\overline{r}^2} + \frac{2}{\overline{r}}\frac{\partial\overline{u}}{\partial\overline{r}} - \frac{2\overline{u}}{\overline{r}^2} = \frac{(1+v)(1-2v)}{(1-v)}\Bigg(&\frac{2\overline{f}}{\overline{c}^2}\frac{\partial\overline{c}}{\partial\overline{r}}\frac{\partial^2\overline{c}}{\partial\overline{r}^2} + \frac{2\overline{f}}{\overline{c}^2}\left(\frac{\partial\overline{c}}{\partial\overline{r}}\right)^2\frac{\partial\overline{\sigma}_h}{\partial\overline{r}} \\ &+ 2\overline{f}\frac{\partial\overline{\sigma}_h}{\partial\overline{r}}\frac{\partial^2\overline{\sigma}_h}{\partial\overline{r}^2} + \frac{2\overline{f}}{\overline{r}\overline{c}^2}\left(\frac{\partial\overline{c}}{\partial\overline{r}}\right)^2 + \frac{2\overline{f}}{\overline{r}}\left(\frac{\partial\overline{\sigma}_h}{\partial\overline{r}}\right)^2 - \frac{2\overline{f}}{\overline{c}^3}\left(\frac{\partial\overline{c}}{\partial\overline{r}}\right)^3 \\ &- \frac{2\overline{f}}{\overline{c}}\frac{\partial^2\overline{c}}{\partial\overline{r}^2}\frac{\partial\overline{\sigma}_h}{\partial\overline{r}} - \frac{2\overline{f}}{\overline{c}}\frac{\partial\overline{c}}{\partial\overline{r}}\frac{\partial^2\overline{\sigma}_h}{\partial\overline{r}^2} - \frac{4\overline{f}}{\overline{r}\overline{c}}\frac{\partial\overline{c}}{\partial\overline{r}}\frac{\partial\overline{\sigma}_h}{\partial\overline{r}}\Bigg) + \frac{(1+v)\,\overline{q}}{3-3v}\frac{\partial\overline{c}}{\partial\overline{r}}, \end{aligned} \tag{2.17}$$

$$\begin{aligned} \frac{\overline{\tau}}{2}\frac{\partial^2\overline{c}}{\partial\overline{t}^2} + \frac{\partial\overline{c}}{\partial\overline{t}} = \frac{\partial^2\overline{c}}{\partial\overline{r}^2} - \frac{\overline{\tau}}{\overline{c}^2}\frac{\partial\overline{c}}{\partial\overline{t}}\left(\frac{\partial\overline{c}}{\partial\overline{r}}\right)^2 + \frac{\overline{\tau}}{\overline{c}}\frac{\partial^2\overline{c}}{\partial\overline{t}\partial\overline{r}}\frac{\partial\overline{c}}{\partial\overline{r}} + \frac{\overline{\tau}}{\overline{c}}\frac{\partial\overline{c}}{\partial\overline{t}}\frac{\partial^2\overline{c}}{\partial\overline{r}^2} - \frac{\partial\overline{c}}{\partial\overline{r}}\frac{\partial\overline{\sigma}_h}{\partial\overline{r}} - \overline{c}\frac{\partial^2\overline{\sigma}_h}{\partial\overline{r}^2} \\ - \overline{\tau}\frac{\partial^2\overline{c}}{\partial\overline{t}\partial\overline{r}}\frac{\partial\overline{\sigma}_h}{\partial\overline{r}} - \overline{\tau}\frac{\partial\overline{c}}{\partial\overline{t}}\frac{\partial^2\overline{\sigma}_h}{\partial\overline{r}^2} + \frac{2}{\overline{r}}\left(\frac{\partial\overline{c}}{\partial\overline{r}} + \frac{\overline{\tau}}{\overline{c}}\frac{\partial\overline{c}}{\partial\overline{t}}\frac{\partial\overline{c}}{\partial\overline{r}} - \overline{c}\frac{\partial\overline{\sigma}_h}{\partial\overline{r}} - \overline{\tau}\frac{\partial\overline{c}}{\partial\overline{t}}\frac{\partial\overline{\sigma}_h}{\partial\overline{r}}\right). \end{aligned} \tag{2.18}$$

The corresponding initial and boundary conditions are

$$\begin{aligned} &\overline{u}|_{\overline{r}=0} = 0, && \overline{\sigma}_r - \overline{\sigma}^V|_{\overline{r}=1} = 0, && \overline{J}|_{\overline{r}=0} = 0, \\ &\overline{c}|_{\overline{r}=1} = 1, && \overline{c}|_{\overline{t}=0} = \frac{31}{33}, && \left.\frac{\partial\overline{c}}{\partial\overline{t}}\right|_{\overline{t}=0} = 0. \end{aligned} \tag{2.19}$$

3. Results and discussion

The primary computational methods for mechanical-chemical coupling problems are the finite element method (Chen *et al.*, 2017), phase field modeling (Chen, 2002), and using COMSOL software (Li *et al.*, 2024). In this paper, we use COMSOL Multiphysics field coupling simulation software to compute the partial differential equations directly. The partial differential equation (PDE) module in COMSOL Multiphysics was used to conduct the simulation study. The mesh was processed using extreme refinement, dividing it into 100 cells with a maximum size

of 1.5×10^{-9}. A transient study was added, with output time steps set at 0.1 s and a total calculation duration of 6 s. The convergence tolerance was adjusted to 1×10^{-7} to ensure simulation accuracy, and the simulation was executed to obtain the numerical results. Table 1 shows the parameter settings for the simulation calculations.

Table 1. Material parameters.

Parameter	Symbol	Values
Elastic modulus [GPa]	E	10
Diffusion coefficient [m^2/s]	D	6.8×10^{-16}
Partial molar volume [m^3/mol]	Ω	3.497×10^{-6}
Poisson's ratio	ν	0.27
Particle radius [m]	r_0	1.5×10^{-7}
Initial concentration [mol/m^3]	c_0	310
Boundary concentration [mol/m^3]	c_1	330
Gas constant [$J/(mol \cdot K)$]	R	8.314
Absolute temperature [K]	T	300
Characteristic time [s]	t_c	0.6
Density [kg/m^3]	ρ	2000

To compare the changes in concentration and stress during the diffusion process under different diffusion forms, a coefficient k_s is introduced in front of the additional terms in Eqs. (2.11) and (2.13). The modified equations are

$$\frac{\partial\left(\sigma_r - k_s\sigma^V\right)}{\partial r} + \frac{2}{r}\left(\sigma_r - k_s\sigma^V - \sigma_\theta\right) = 0, \qquad J = \left(c + k_s t_c \frac{\partial c}{\partial t}\right) v_k, \tag{3.1}$$

k_s is 1 for the theoretical model used in this paper and k_s is 0 for the theoretical model used for non-Fick diffusion, the difference between the two theories can be clearly seen through the equation. If no special instructions are given, k_s takes 1.

Figure 2 presents a comparison of concentration, displacement, radial stress, and hoop stress between the non-Fick diffusion theory (non-Fick, $k_s = 0$) and the mechano-diffusion coupling model proposed in this section (our model, $k_s = 1$) at a characteristic time value of $\overline{\tau} = 1.8 \times 10^{-2}$ ($t_c = 0.6$ s). Ions are seen to progressively spread from the edge towards the center in Fig. 2a. When $k_s = 0$ and $k_s = 1$, the concentration changes abruptly before diffusion reaches the sphere's center. There are no abrupt variations in concentration for either model after diffusion reaches the sphere's center ($\overline{t} = 0.18$), and the concentration variations along the radial direction are similar. At the same position, the ion concentration increases with increasing diffusion time, and the ion concentration under our model is higher than that under the non-Fick diffusion model. This is because the presence of characteristic time in our model leads to an increase in the ion diffusion flux, which in turn results in a higher ion concentration over a specific period of time. The spherical model's displacement diagram is shown in Fig. 2b. The graphic demonstrates that ion diffusion causes displacement within the sphere. At the same moment, the radial displacement in our model is greater than that in the non-Fick diffusion model due to ion diffusion.

It is evident from Fig. 2c that the radial stress rises with proximity to the sphere's center. In our model, the radial stress in the region reached by ion diffusion is lower than that under non-Fick diffusion at the same time and location. This is because, within the same time frame, the increase in diffusion flux leads to a higher ion concentration and greater displacement within our model. According to the formula for radial stress, an increase in displacement leads to an increase in radial stress, while an increase in concentration leads to a decrease in radial

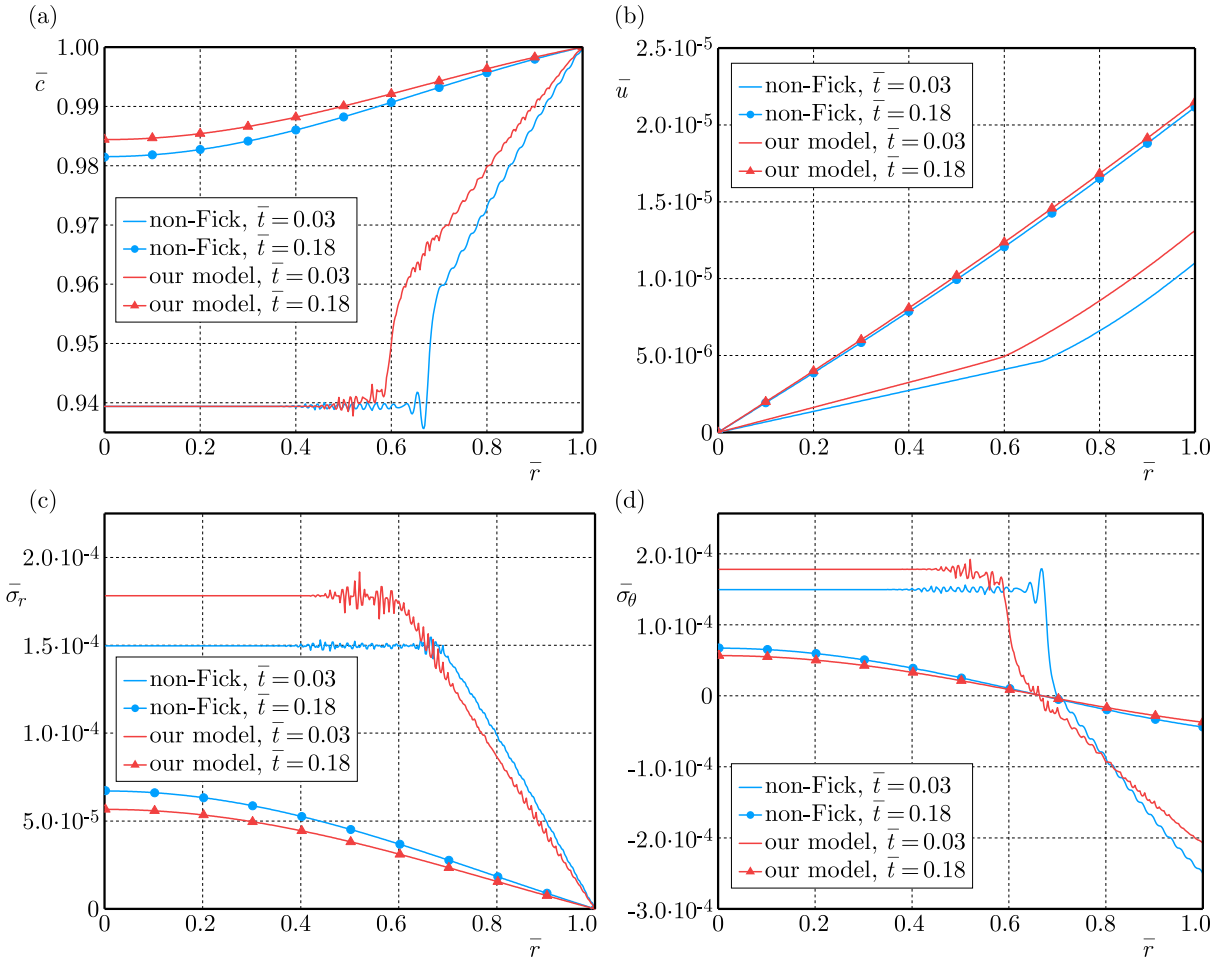

Fig. 2. Comparison of concentration (a); displacement (b); radial stress (c); and hoop stress (d) between non-Fick diffusion and our model.

stress. However, the effect of displacement on radial stress is less significant than the effect of concentration on radial stress. Therefore, in our mechanic-diffusion coupling model, the radial stress is less than that in the non-Fick mechano-diffusion coupling model. In regions where ions have not diffused, the concentration remains unchanged, and radial stress is influenced solely by the changes in displacement. Since both the displacement and its gradient are greater in our model compared to the non-Fick diffusion model, in areas where the concentration has not changed at the same moment, the radial stress in our model is greater than that in the non-Fick diffusion model. From Fig. 2d, it can be observed that in both scenarios, the hoop stress is compressive near the outer surface and tensile near the center of the sphere. As the ion moves from the outer surface towards the center, the magnitude of the hoop stress first decreases to zero and then gradually increases. In the regions where ions have not diffused, the hoop stress in our model is greater than that in the non-Fick diffusion model. At $\bar{t} = 0.18$, within the sphere at the same moment, the hoop stress in our model is less than that under both non-Fick and Fick diffusion, for the same reasons as with the radial stress.

Figure 3 illustrates the distribution of ion concentration, displacement, radial stress, and hoop stress within the model at different times for a characteristic time of $\overline{\tau} = 1.8 \times 10^{-2}$ (0.6 s). From Fig. 3a, it can be observed that during the diffusion process, ions diffuse from the outer surface of the sphere towards the center. Moreover, as indicated by the blue, green, and red curves, there is no change in ion concentration near the center during the period when ions have not yet diffused to that area. This is due to the fact that in our model, the diffusion speed of ions is not infinite; ions diffuse at a finite speed. When concentration boundary conditions

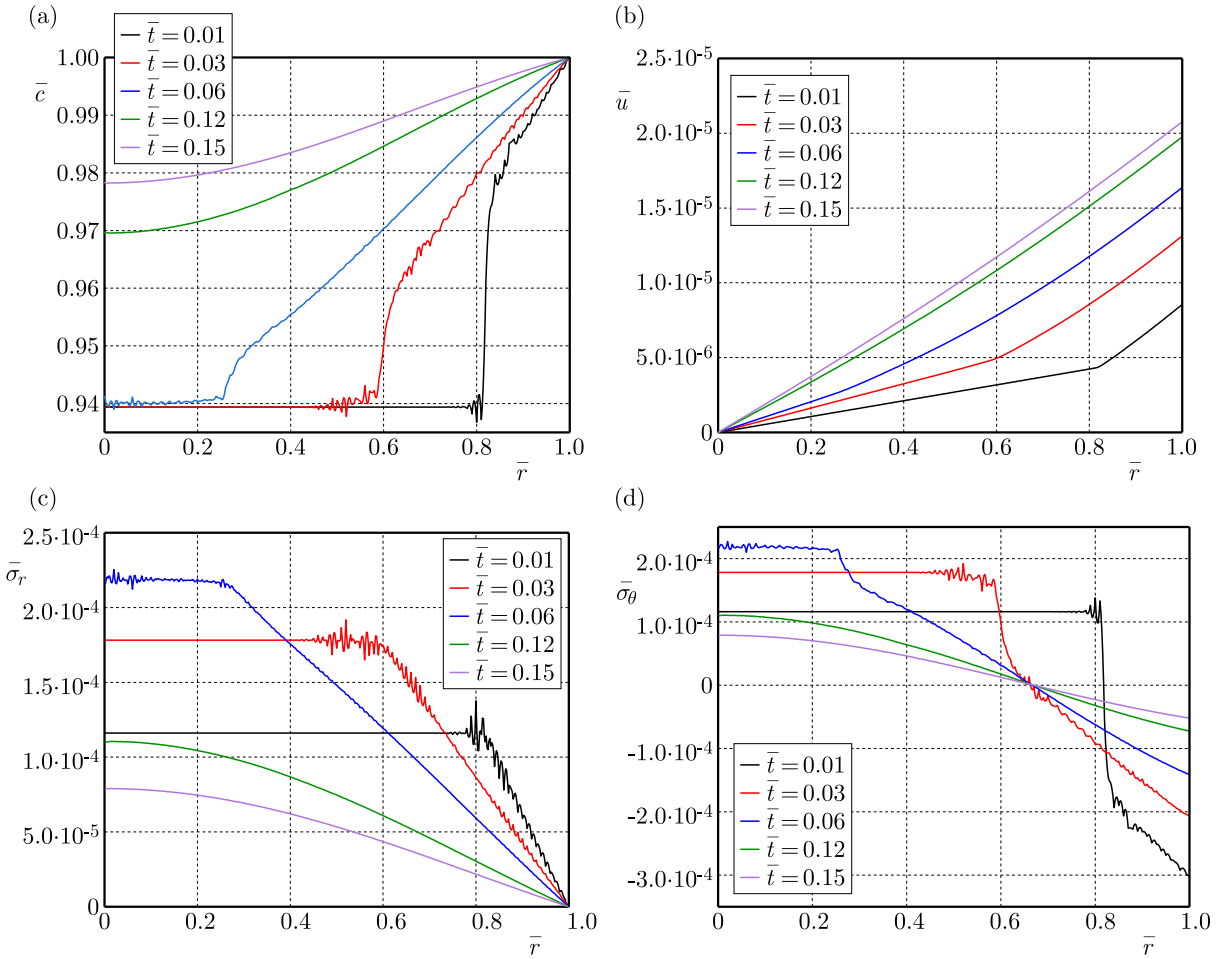

Fig. 3. Distribution of concentration (a); displacement (b); radial stress (c); and hoop stress (d) at different diffusion times.

are applied, the ion concentration within the model does not change immediately but senses the change after a specific period. Furthermore, within a specific time frame, the influence of the second-order terms causes the concentration diffusion curves to exhibit oscillatory behavior. After a period of time, as the impact of the second-order terms diminishes, the oscillatory nature of the curves disappears. It is evident from Fig. 3b, which displays the displacement, that the model's displacement rises as the diffusion time does. Before the diffuse reaches the center of the sphere, at three time points $\bar{t} = 0.01$, $\bar{t} = 0.03$, and $\bar{t} = 0.06$, it can be seen that in the region where the concentration has not changed, the model's displacement maintains a linear change. This also confirms the reason why the radial stress and hoop stress in the region where the concentration has not changed remain constant.

It is evident from Fig. 3c that the model's radial stress is tensile and increases gradually from the outer surface towards the center of the sphere in the region where ions have diffused. When $\bar{t} = 0.01$, $\bar{t} = 0.03$, and $\bar{t} = 0.06$, in the regions where ions have not diffused, the radial stress remains constant, while in the regions where ions have diffused, the radial stress shows a decreasing trend with increasing time. According to the formula, the radial stress is determined by the ion concentration and displacement. An increase in concentration leads to a decrease in radial stress, while an increase in displacement leads to an increase in radial stress. In the region where ions have diffused, as the diffusion time increases, the concentration change at the same position becomes greater. The effect of concentration change on radial stress is more significant than that of displacement change. Therefore, in the region where ions have diffused, the radial stress decreases with increasing time. In the regions where ions have not diffused (the platform

parts of the red, blue, and black solid lines), the concentration remains unchanged, and the radial stress is only affected by the changes in displacement. At this time, the radial stress decreases with the increase in time. It is shown by Fig. 3d that the hoop stress of the model is compressive at the outer surface and tensile at the center of the sphere. In the region where ions have diffused, the hoop stress decreases from the outer surface towards the center, reaching zero before gradually increasing. The hoop stress also decreases over time. From Fig. 3d, at the time points $\bar{t} = 0.01$, $\bar{t} = 0.03$, and $\bar{t} = 0.06$, it can be seen that in the regions where ions have not diffused, the hoop stress remains unchanged.

From Figs. 4a and 4b, it can be observed that the ion concentration and displacement within the model increase with the increase in characteristic time. The characteristic time affects the ion diffusion flux, which in turn increases with the characteristic time. Therefore, within the same time frame, the higher the characteristic time, the higher the ion concentration.

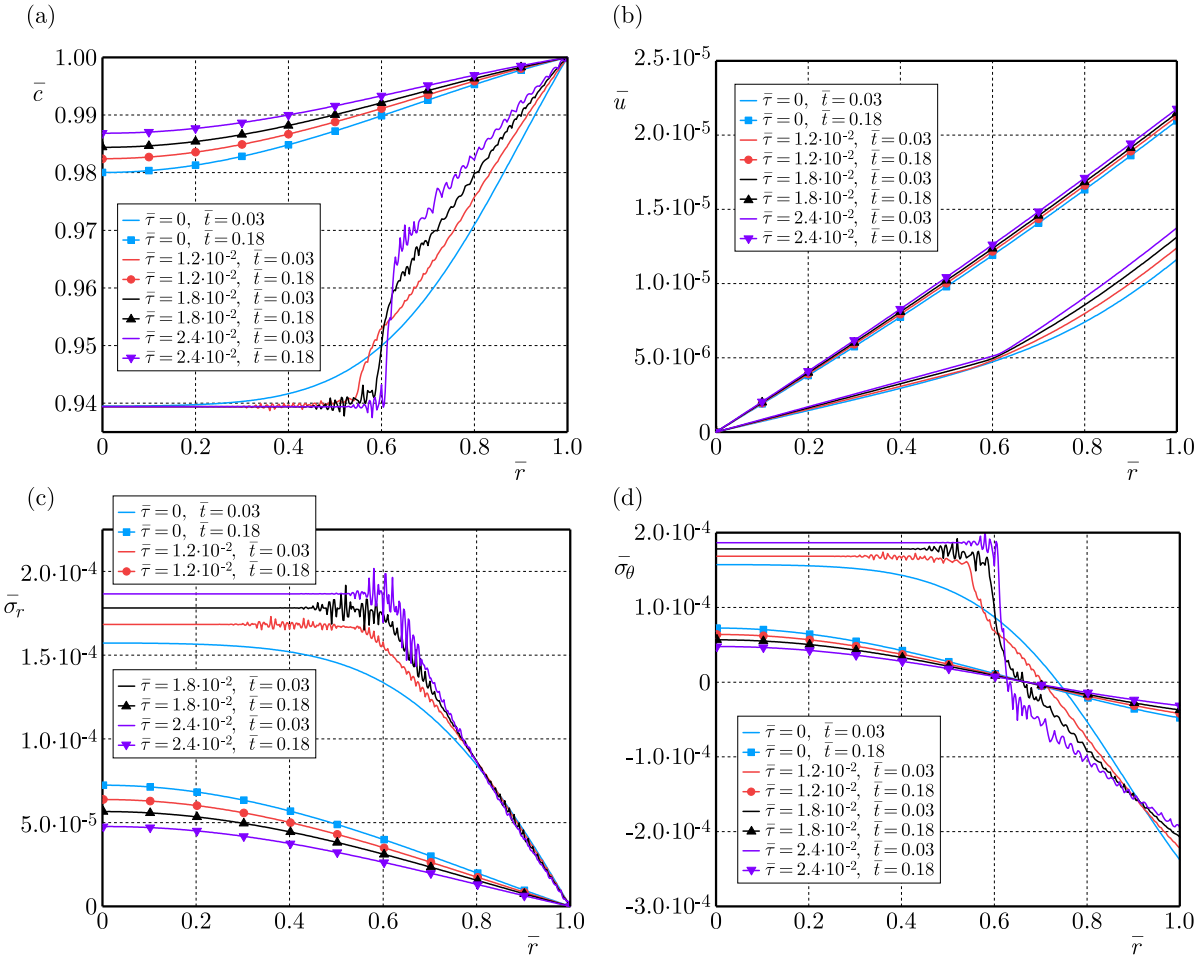

Fig. 4. Distribution of concentration (a); displacement (b); radial stress (c); and hoop stress (d) under different characteristic times.

According to Figs. 4c and 4d, at $\bar{t} = 0.03$ ($t = 1$ s, no graphic line) in the regions where ions have not diffused, both radial stress and hoop stress increase with the increase in characteristic time. This is because the displacement within the model increases with characteristic time, and when the concentration remains unchanged, both radial stress and hoop stress increase with the increase in displacement. After the ions have diffused to the center of the sphere ($\bar{t} = 0.18$, graphic line), both radial stress and hoop stress within the model decrease with the increase in characteristic time. This is because both radial stress and hoop stress are determined by changes in displacement and concentration, with the impact of concentration changes on radial and hoop stress being greater than the effects produced by changes in displacement. As the concentration

change increases with the increase in characteristic time, in the curve at $\bar{t} = 0.18$, the radial stress and hoop stress decrease with the increase in characteristic time.

4. Conclusion

Based on the theory of transient mechano-chemical coupling, this paper investigates the mechano-chemical coupled diffusion processes of ions within a spherical structure, exploring the effects of characteristic time and boundary concentration on the distribution of concentration, displacement, radial stress, and hoop stress within the structure. The results indicate that:

1) The presence of characteristic time transforms the classical chemical diffusion control equation from a parabolic type to a hyperbolic type, changing the diffusion speed from infinite to finite. Concentration curves exhibit oscillatory behavior. Consequently, the concentration, stress, and displacement curves distinctly show two regions: the area that has been diffused and the area that has not been diffused. At the interface between these two regions, there is a sudden change in the concentration and stress curves.
2) The characteristic time influences the magnitude of ion concentration, displacement, radial stress, and hoop stress. Ion concentration and displacement increase with the increase in characteristic time. Radial stress and hoop stress increase in regions where ions have not yet diffused with the increase in characteristic time, while they decrease in regions where ions have already diffused with the increase in characteristic time.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant no. 12272095 and 12402212).

References

1. Chen, J.Y., Wang, H.L., Yu, P.F., & Shen, S.P. (2017). A finite element implementation of a fully coupled mechanical-chemical theory. *International Journal of Applied Mechanics*, *9*(3), Article 1750040. https://doi.org/10.1142/s1758825117500405
2. Chen, L.Q. (2002). Phase-field models for microstructure evolution. *Annual Review of Materials Research*, *32*, 113–140. https://doi.org/10.1146/annurev.matsci.32.112001.132041
3. Chen, X.Q., Deng, F., & Shen, S.P. (2023). Chemomechanical finite element analysis for surface oxidation of Aluminum alloy. *Acta Mechanica*, *234*(4), 1713-1732. https://doi.org/10.1007/s00707-022-03463-5
4. Chu, J.L., & Lee, S.B. (1994). The effect of chemical stress on diffusion. *Journal of Applied Physics*, *75*(6), 2823–2829. https://doi.org/10.1063/1.356174
5. Dai, L., & Xiao, R. (2021). A thermodynamic-consistent model for the thermo-chemo-mechanical couplings in amorphous shape-memory polymers. *International Journal of Applied Mechanics*, *13*(2), Article 2150022. https://doi.org/10.1142/s1758825121500228
6. Dong, H.X., & Jiang, R.Q. (1995). Theoretical study of mass transfer behavior in transient mass transfer process (in Chinese). *Journal of Harbin Engineering University*, *16*(3), 88–94.
7. Hu, H., Yu, P.F., & Suo, Y.H. (2020). Stress induced by diffusion and local chemical reaction in spherical composition-gradient electrodes. *Acta Mechanica*, *231*(7), 2669–2678. https://doi.org/10.1007/s00707-020-02652-4
8. Hu, S., & Shen, S. (2013). Non-equilibrium thermodynamics and variational principles for fully coupled thermal-mechanical-chemical processes. *Acta Mechanica*, *224*(12), 2895–2910. https://doi.org/10.1007/s00707-013-0907-1

9. Konica, S., & Sain, T. (2020). A thermodynamically consistent chemo-mechanically coupled large deformation model for polymer oxidation. *Journal of the Mechanics and Physics of Solids*, *137*, Article 103858. https://doi.org/10.1016/j.jmps.2019.103858

10. Kuang, Z.B. (2014). Discussions on the temperature wave equation. *International Journal of Heat and Mass Transfer*, *71*, 424–430. https://doi.org/10.1016/j.ijheatmasstransfer.2013.12.016

11. Li, J.C.M. (1978). Physical chemistry of some microstructural phenomena. *Metallurgical and Materials Transactions A*, *9*(10), 1353–1380. https://doi.org/10.1007/BF02661808

12. Li, X., Gao, J., Zhou, W., & Li, H. (2024). Application of COMSOL multiphysics in lithium-ion batteries (in Chinese). *Energy Storage Science and Technology*, *13*(2), 546–567.

13. Liu, K.C. (2007). Simultaneous estimation of relaxation time and diffusion coefficient during the rapid transient mass transfer. *Journal of the Chinese Society of Mechanical Engineers*, *28*(5), 541–547. https://doi.org/10.29979/JCSME.200710.0009

14. Nguyen, T.T., Waldmann, D., & Bui, T.Q. (2019). Computational chemo-thermo-mechanical coupling phase-field model for complex fracture induced by early-age shrinkage and hydration heat in cement-based materials. *Computer Methods in Applied Mechanics and Engineering*, *348*, 1–28. https://doi.org/10.1016/j.cma.2019.01.012

15. Shen, S.P. (2022). Transient continuum mechanics and chemomechanics. *Journal of Applied Mechanics*, *89*(6), Article 061004. https://doi.org/10.1115/1.4054061

16. Suo, Y.H., Hu, H., & Liu, J. (2021). Fully chemo-mechanical coupling analysis of a spherical electrode with reversible chemical reaction. *International Journal of Energy Research*, *45*(6), 9667–9676. https://doi.org/10.1002/er.6430

17. Suo, Y.H., & Shen, S.P. (2013a). Analytical solution for 2D non-Fickian transient mass transfer with arbitrary initial and periodic boundary conditions. *ASME Journal of Heat and Mass Transfer*, *135*(8), Article 082001. https://doi.org/10.1115/1.4024352

18. Suo, Y.H., & Shen, S.P. (2013b). Analytical solution for one-dimensional coupled non-Fick diffusion and mechanics. *Archive of Applied Mechanics*, *83*(3), 397–411. https://doi.org/10.1007/s00419-012-0687-4

19. Suo, Y.H., Yang, H., & Jia, Q.N. (2024). Coupled diffusion-mechanical analysis with dislocation effect in porous spherical electrode. *Solid State Ionics*, *404*, Article 116422. https://doi.org/10.1016/j.ssi.2023.116422

20. Yan, G.S., Wu, Y.H., Liu, W.Y., Yu, W.S., & Shen, S.P. (2024). Stress evolution in enamel coating/Ni-based alloy systems during isothermal oxidation. *Journal of Applied Physics*, *135*(8), Article 085103. https://doi.org/10.1063/5.0185084

21. Yu, P.F., Hu, S.L., & Shen, S.P. (2016). Electrochemomechanics with flexoelectricity and modelling of electrochemical strain microscopy in mixed ionic-electronic conductors. *Journal of Applied Physics*, *120*(6), Article 065102. https://doi.org/10.1063/1.4960445

22. Yu, P.F., & Shen, S.P. (2014). A fully coupled theory and variational principle for thermal-electrical-chemical-mechanical processes. *Journal of Applied Mechanics*, *81*(11), Article 111005. https://doi.org/10.1115/1.4028529

23. Zhang, K., Li, Y., Wang, F., Zheng, B.L., & Yang, F.Q. (2019). A phase-field study of the effect of local deformation velocity on lithiation-induced stress in wire-like structures. *Journal of Physics D: Applied Physics*, *52*(14), Article 145501. https://doi.org/10.1088/1361-6463/ab00dc

Manuscript received February 18, 2025; accepted for publication June 9, 2025; published online October 17, 2025.

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 889–902, Warsaw 2025
https://doi.org/10.15632/jtam-pl/207524

EXPERIMENTAL INVESTIGATION OF THE LONG-TERM MECHANICAL BEHAVIOR OF MUDSTONE UNDER VARYING WATER CONTENTS

Zhuangen QIN

China Anneng Group Third Engineering Bureau Co.Ltd., Chengdu, China
qinzexx@sina.com

This study presents a series of uniaxial compression and creep tests designed to elucidate the long-term mechanical properties of mudstone subjected to different water content conditions. The results demonstrate that water content exerts a significant influence on the short-term strength, elastic modulus, and creep response of mudstone. Specifically, the uniaxial compressive strength and elastic modulus exhibit an exponential decrease with increasing water content. Furthermore, the creep behavior of mudstone is markedly affected by water content. A creep damage model, integrating the Burgers model with water-induced and creep-induced damage variables, is proposed.

Keywords: mudstone; water content; creep; damage variable; creep damage model.

1. Introduction

Mudstone, a widely distributed sedimentary rock, plays a vital role in geotechnical engineering applications, including slope stabilization, tunneling, and foundation design. Its mechanical behavior is notably sensitive to environmental factors, particularly water content. Variations in water content can induce substantial alterations in mudstone's strength, deformability, and long-term stability. For instance, in slope engineering, cyclical wetting-drying processes resulting from rainfall or groundwater fluctuations may trigger swelling, softening, and potentially catastrophic failure of mudstone slopes (Qi *et al.*, 2024; Yu *et al.*, 2024). Similarly, in underground excavations, water infiltration can accelerate creep deformation, thereby jeopardizing structural integrity (Li *et al.*, 2023; Wang *et al.*, 2021). This is especially relevant in areas prone to landslides, such as the Three Gorges Reservoir Area in China, where the purple mudstone of the Middle Triassic Badong Formation is particularly susceptible to creep-related failures (Wang *et al.*, 2021). Despite its practical relevance, the long-term mechanical response of mudstone under variable water conditions remains incompletely understood, particularly in scenarios involving sustained loads, such as dam foundations or deep tunnels. A thorough understanding of these mechanisms is essential for optimizing engineering designs, mitigating risks, and ensuring the long-term safety of infrastructure projects. This study addresses this knowledge gap by systematically investigating the influence of water content on both the short-term and long-term mechanical properties of mudstone.

Numerous investigations have explored the relationship between water content and the mechanical behavior of mudstone (Chen *et al.*, 2022; Liu *et al.*, 2021; Shao *et al.*, 2024; Zheng *et al.*, 2024). These studies have highlighted the critical role of water in weakening mudstone's mechanical properties. For instance, several studies have focused on the impact of water content on the unconfined compressive strength (UCS) and Young's modulus, demonstrating a marked decrease in these parameters with increasing water (Chen *et al.*, 2022; Gao *et al.*, 2023). Microstructural analyses, using techniques like scanning electron microscopy (SEM), have further revealed that

water intrusion leads to changes in cementation, compaction, and pore size distribution within the mudstone matrix, ultimately contributing to the degradation of its mechanical properties (Gao *et al.*, 2023; Shao *et al.*, 2024; Zheng *et al.*, 2024).

The time-dependent behavior of mudstone under varying water conditions has also been a subject of extensive research. Creep tests conducted under different confining pressures and water contents have demonstrated that water significantly alters the creep characteristics of mudstone (Li *et al.*, 2023; Wang *et al.*, 2021). Sawatsubashi *et al.* (2021) found that immersion-induced creep deformation is influenced by both initial water content and shear stress, with higher initial water content leading to increased creep deformation. Moreover, cyclic wetting-drying tests have shown that repeated cycles exacerbate the deterioration of mudstone's mechanical properties, leading to a transition from brittle to ductile behavior and a reduction in shear strength (Yu *et al.*, 2024; Yang *et al.*, 2022). The damage mechanisms associated with these cycles involve swelling and shrinkage of clay minerals, dissolution of soluble minerals, and the development of microcracks (Yang *et al.*, 2022; Yu *et al.*, 2024).

Constitutive models have been developed to capture the complex behavior of mudstone under varying water and stress conditions. Wang *et al.* (2020) proposed a nonlinear disturbance creep damage model based on the Burgers model to account for the influence of cyclic disturbance loads on mudstone creep. Ma *et al.* (2018) developed a new shear rheological model to describe the behavior of soft interlayers with varying water content. Other researchers have focused on developing damage constitutive models that incorporate the effects of rock-water interactions and cyclic wetting-drying (Yu *et al.*, 2024). Liu *et al.* (2024) developed a time-dependent expansion model for mudstone submerged in water, based on rheological theory, while Ping *et al.* (2024) constructed a mechanical damage model to predict the behavior of gypsum-bearing mudstone under varying dissolution times. These models highlight the importance of considering both the instantaneous and time-dependent effects of water on the mechanical behavior of mudstone. Moreover, Yang *et al.* (2019) investigated deformation and failure of mudstone under triaxial compression using experiment and particle flow code, which can be significant for the design of deep tunnel support. Additionally, recent studies by Shao *et al.* (2024) and Zheng *et al.* (2024) have explored the influence of microstructure on the mechanical behavior of similar geomaterials, providing valuable insights into the damage mechanisms at play.

However, existing studies exhibit several limitations. First, the long-term creep behavior under multi-stage stress conditions, particularly in the presence of varying water content, requires further investigation. Second, the interaction between water content and confining pressure in controlling creep damage has not been fully explored. Third, existing constitutive models often simplify the damage evolution process and fail to comprehensively consider the microstructural degradation caused by water. These limitations hinder the accurate prediction of the long-term performance of mudstone in practical engineering applications. This study aims to systematically investigate the influence of water content on the short-term strength, elastic modulus, and creep behavior of mudstone, and to develop a creep damage model that incorporates the effects of water content.

2. Materials and method

2.1. Specimen preparation

The mudstone specimens used in this study were obtained from a slope excavation site. This site is located in a geological environment characterized by sedimentary rocks, and the mudstone specimens are representative of the local geological stratum. The mudstone samples present as maroon-colored, dense blocks. The surface exhibits a soil-like texture, and microscopic observation reveals a silt-like structure. The average unit weight of the natural rock is approximately $2260\,\mathrm{kg/m^3}$, indicating a relatively high density compared to other sedimentary rock types.

All specimens were prepared as standard cylindrical specimens, adhering strictly to the guidelines of the International Society for Rock Mechanics (Bieniawski & Bernede, 1979). Initially, the mudstone blocks were cut into approximate cylindrical shapes using a diamond-wire saw. Subsequently, the specimens were polished using a grinding machine to ensure dimensional precision. The final specimens had a diameter of 50 mm and a height of 100 mm. To ensure the accuracy of the experimental results, the flatness of the end surfaces of each specimen was controlled to within 0.03 mm. This high-precision preparation minimizes the influence of specimen geometry and surface roughness on the experimental results, thus ensuring the reliability of the subsequent mechanical property tests. The detailed parameters of each specimen are shown in Table 1.

Table 1. Specimens parameters.

Specimen number	Water content [%]	Diameter [mm]	Height [mm]	Weight [g]	Density [g/cm^3]
uc01	0.00	50.01	100.04	443.68	2.26
uc02	0.26	49.69	99.77	442.53	2.29
uc03	0.72	49.92	100.08	441.44	2.25
uc04	1.64	50.02	100.16	442.32	2.25
cr-01	0.00	49.87	100.24	445.44	2.28
cr-02	0.26	49.88	100.46	442.71	2.26
cr-03	0.72	49.75	102.42	442.72	2.22
cr-04	1.64	50.01	99.79	445.52	2.27

2.2. Determination of water content

The water content of the mudstone specimens was controlled by oven drying and vacuum saturation methods. First, the natural mudstone specimens were weighed and then placed in an oven at 105 °C for 24 hours for complete drying, after which they were weighed again. Subsequently, portions of the dried specimens were vacuum-saturated in a vacuum saturator for a minimum of 12 hours to determine the saturated water content. Additional specimens were immersed in deionized water for 3 hours and 6 hours to achieve intermediate water contents. In this experiment, four distinct water content levels were employed: 0 % (dry), 0.26 %, 0.72 %, and 1.64 % (saturated), as indicated in Table 1.

2.3. Testing apparatus and procedure

2.3.1. Testing apparatus

The uniaxial compression tests were performed using a TFD-2000 electro-hydraulic servo rock triaxial testing machine (Fig. 1). This machine is capable of acquiring high- and low-speed data with excellent dynamic response, static stability, and system stiffness. The maximum axial load capacity is 2000 kN, and the maximum confining pressure and pore pressure are 100 MPa, with an accuracy of 0.1 %. The system comprises an axial loading system, a fluid pressure loading system (for water and confining pressure), a data acquisition system, and a central control system. The machine is capable of performing uniaxial compression, triaxial compression, seepage, creep, direct tensile, and indirect tensile tests, and it can synchronously record stress-strain curves throughout the rock loading process.

2.3.2. Testing procedure

A series of step-loading creep tests were conducted on mudstone specimens with different water contents, as summarized in Table 2. The load levels for each step were determined as 20 %, 35 %, 50 %, 65 %, and 80 % of the uniaxial compressive strength of mudstone at the corresponding

Fig. 1. TFD-2000 electro-hydraulic servo rock triaxial testing machine.

Table 2. Step-loading values for creep tests.

Specimen number	Water content [%]	σ_c [MPa]	1st stage load [MPa]	2nd stage load [MPa]	3rd stage load [MPa]	4th stage load [MPa]	5th stage load [MPa]
cr-01	0.00	1.43	0.286	0.50	0.72	0.93	1.14
cr-02	0.26	0.77	0.154	0.27	0.39	0.50	0.62
cr-03	0.72	0.56	0.112	0.20	0.28	0.36	
cr-04	1.64	0.44	0.088	0.15	0.22		

water content. During the test, each load level was applied at a loading rate of 0.5 MPa/s until reaching the designated value, and then maintained for 48 hours.

After the 48-hour creep period at each load level, the specimens were carefully examined. If no signs of failure, such as a significant increase in the deformation rate or visible cracks on the specimen surface, were observed, the experiment proceeded to the next loading stage. However, if the specimens exhibited a significant increase in the deformation rate, indicating the onset of the accelerating creep stage and imminent bearing capacity loss, or if visible cracks appeared on the specimen surface, signifying severe damage to the internal structure, the specimens were considered to have experienced creep failure, and the test was terminated. This experimental procedure ensured the accurate investigation of the creep behavior of mudstone under different water content and stress conditions, providing reliable data for subsequent analysis of the long-term mechanical properties of mudstone.

3. Test results and analysis

3.1. Short-term strength

Prior to the long-term experiments, a series of uniaxial compression tests were performed on mudstone specimens with varying water contents. The resulting stress-strain curves are presented in Fig. 2.

A comparison of the post-peak stress-strain curves of specimens with different water contents reveals a distinct influence of water content on the mechanical behavior of mudstone. Under lower water content conditions, the rock specimens exhibit characteristic brittle behavior. Upon reach-

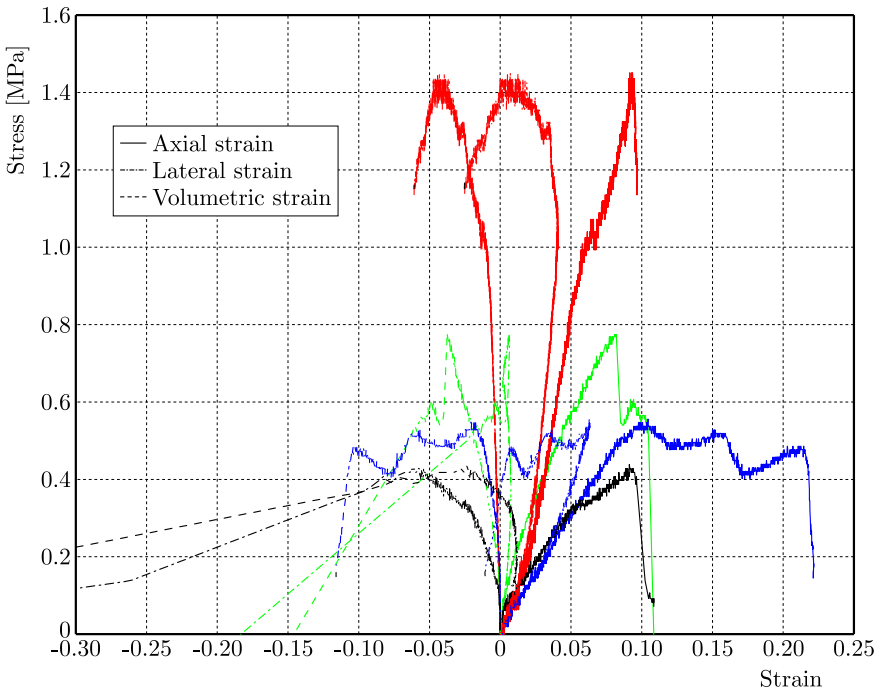

Fig. 2. Stress-strain curves of mudstone under different water content.

ing the peak stress, abrupt failure of the internal structure occurs due to the rapid accumulation of internal stress. The specimen fails suddenly, and the stress drops precipitously, indicating an almost instantaneous loss of bearing capacity. Conversely, under higher water content conditions, the brittleness of the mudstone is reduced, and the behavior becomes more plastic. After reaching the peak stress, the specimen can sustain a certain amount of deformation without immediate failure. This is attributed to the water-lubricated particles' ability to adjust their positions, dissipating the energy of external forces through plastic deformation. Consequently, the stress-strain curve exhibits a more gradual decline, indicating that the specimen retains some capacity to resist further deformation after the peak stress.

Figure 3 illustrates the relationship between uniaxial compressive strength, elastic modulus, and water content for the mudstone specimens. As the water content increases, both the uniaxial

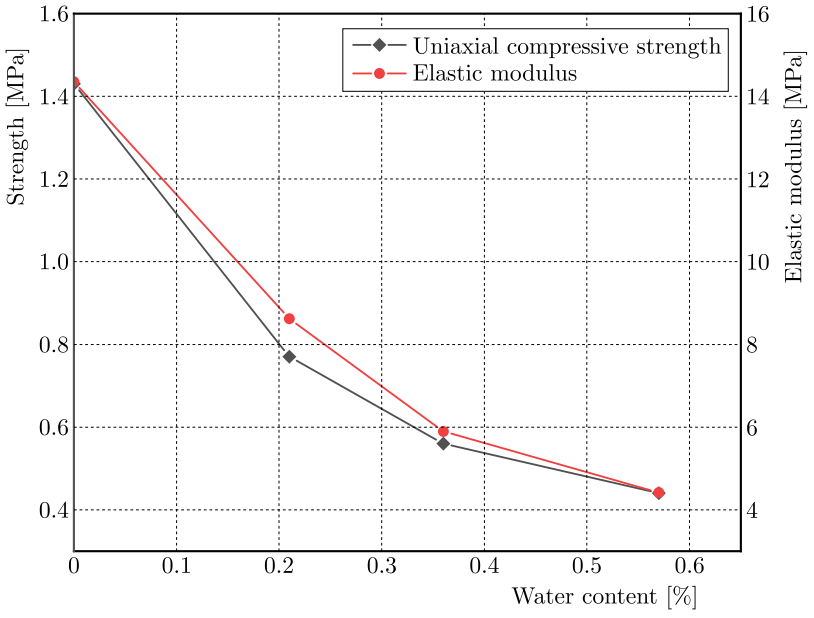

Fig. 3. Uniaxial compressive strength and elastic modulus vs water content.

compressive strength and elastic modulus of the mudstone decrease continuously. The initial uniaxial compressive strength of the dry mudstone specimen was 1.43 MPa, while at a water content of 1.64 % (saturation), the strength decreased to 0.44 MPa, representing a reduction of 69.45 %. Similarly, the elastic modulus decreased significantly with increasing water content, from 14.35 MPa for the dry specimen to 4.12 MPa for the saturated specimen, a reduction of 69.23 %. To quantify the statistical significance of the observed trends, a one-way ANOVA was performed on the uniaxial compressive strength and elastic modulus data. The results indicated a statistically significant effect of water content on both parameters ($p < 0.05$). Post-hoc tests (Tukey's HSD) revealed significant differences between the strength and modulus values at different water content levels.

Statistical analysis of the uniaxial compressive strength and elastic modulus data revealed an exponential decline in both properties with increasing water content. The relationships between these properties and water content were fitted to derive specific expressions. These expressions can be used to predict the uniaxial compressive strength and elastic modulus of mudstone under different water content conditions, providing a valuable tool for engineering design.

3.2. Creep behavior of mudstone under different water content

The uniaxial creep curves of mudstone under different water content conditions are shown in Fig. 4. At low water content, the creep deformation of mudstone is relatively small. The creep rate decreases rapidly during the primary creep stage. This behavior is attributed to the relatively stable internal structure of mudstone at low water content. The particles are tightly bound, and the resistance to deformation is high. The specimen exhibits some elastic-like behavior, and creep deformation is primarily due to elastic-recovery-like adjustments within the internal structure of the mudstone under external stress.

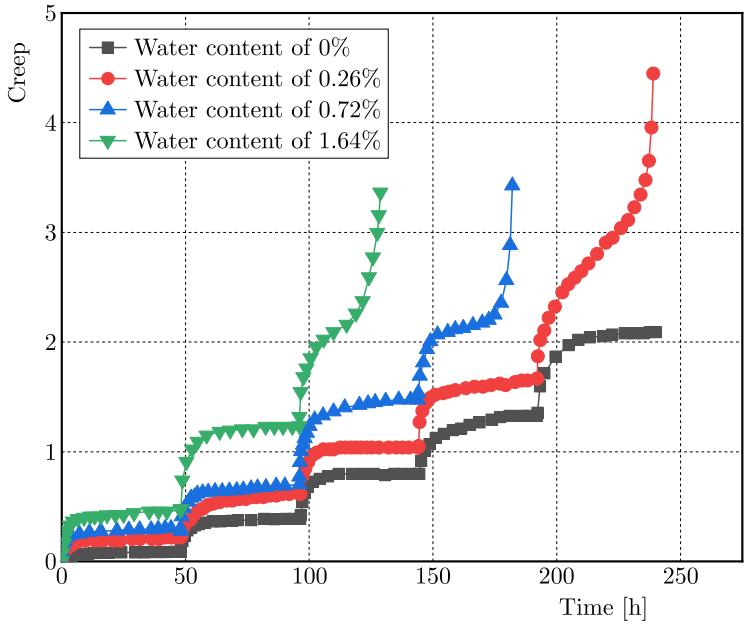

Fig. 4. Uniaxial creep curves of mudstone under different water contents.

As the water content increases, the creep deformation of mudstone increases significantly, and the primary creep stage becomes shorter. This is because the increase in water content leads to the softening of the mudstone. Water in the pores further lubricates the particles, reducing friction between them. The internal structure of the mudstone becomes looser, facilitating particle movement and rearrangement under external stress. As a result, the specimen enters the secondary creep stage earlier. In the secondary creep stage, although the creep rate becomes relatively stable, the overall creep deformation is much greater than that at low water content.

For specimens with different water contents, the creep curves share common characteristics yet also display differences. Figures 5, 6, and 7 all illustrate the three-stage creep process: decelerating (primary) creep, steady-state (secondary) creep, and accelerating (tertiary) creep. However, due to varying water contents and applied loads, their behaviors differ. The specimen with 0.26 % water content under a load of 0.616 MPa, the one with 0.72 % water content under 0.364 MPa, and the 1.64% water-content specimen under 0.22 MPa all start with a decelerating creep stage where the creep rate decreases rapidly as the internal structure adjusts to the load. This is followed by the steady-state creep stage with a relatively stable creep rate. Finally, they enter the accelerating creep stage as the internal structure deteriorates. The specimen with a higher water content generally shows a shorter overall creep-time-to-failure. For example, the 1.64 % water-content specimen reaches the accelerated creep stage and fails more quickly

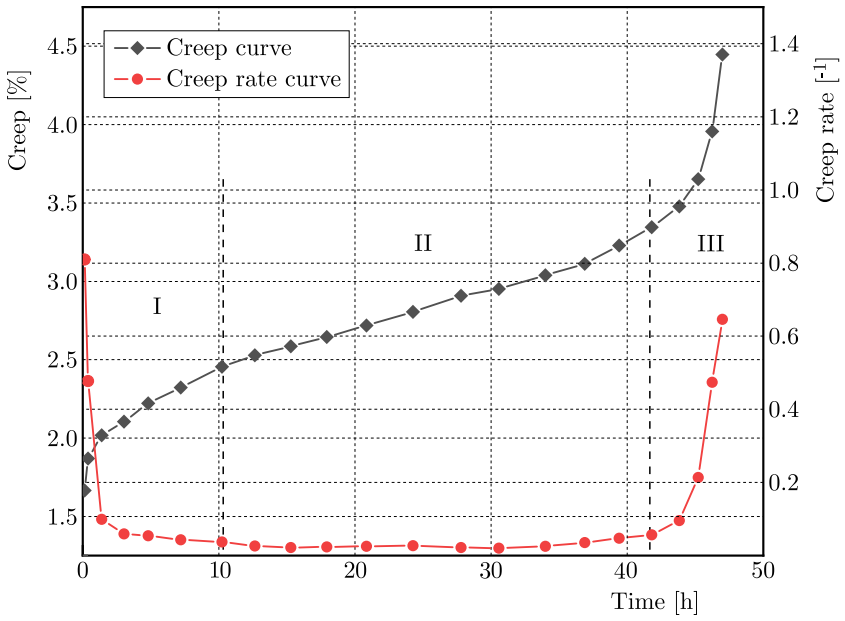

Fig. 5. Uniaxial creep and creep rate curves of mudstone with 0.26 % water content under the fifth-stage load of 0.616 MPa.

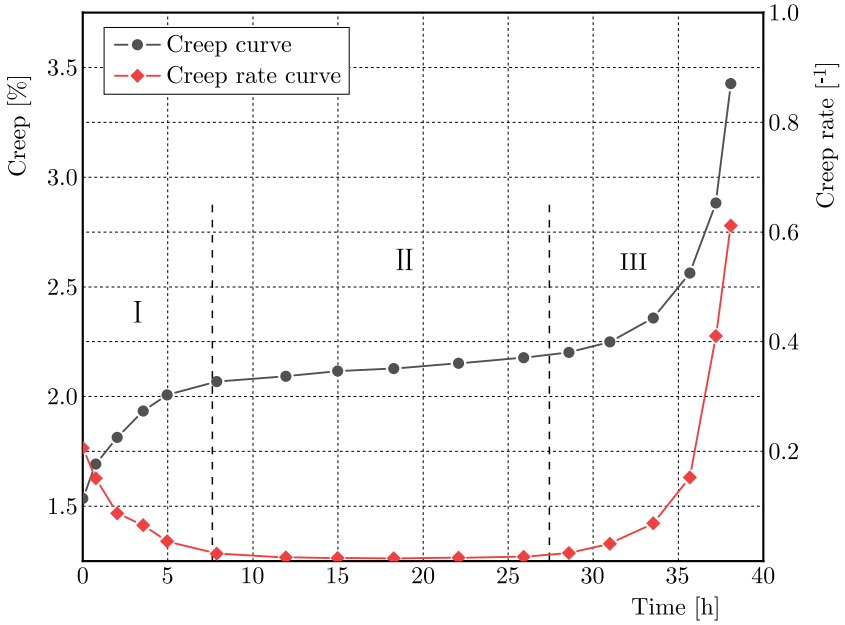

Fig. 6. Uniaxial creep and creep rate curves of mudstone with 0.72 % water content under the fourth-stage load of 0.364 MPa.

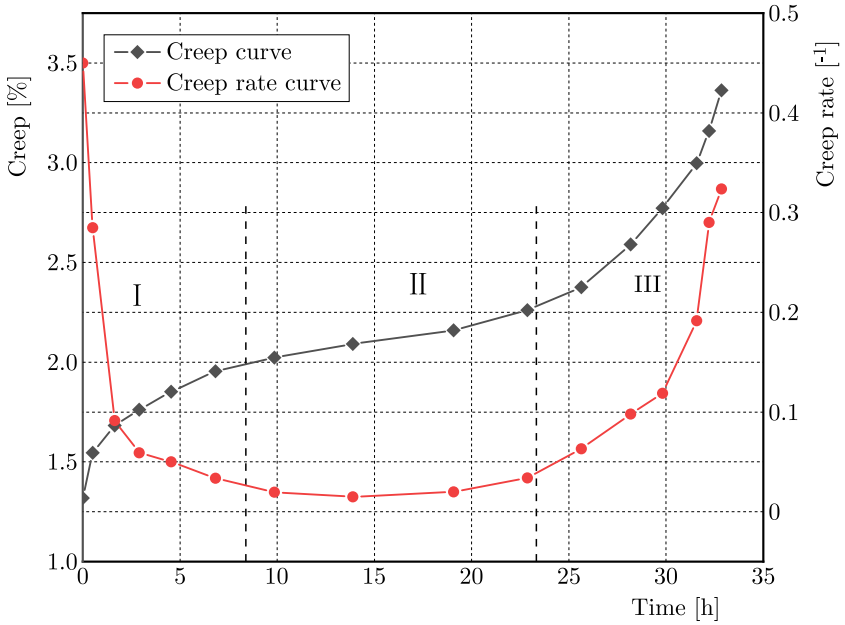

Fig. 7. Uniaxial creep and creep rate curves of mudstone with 1.64 % water content under the third-stage load of 0.22 MPa.

compared to the 0.26 % and 0.72 % water-content specimens, indicating that water content significantly influences the creep failure time and overall creep behavior of mudstone.

4. Creep damage constitutive model

4.1. Burgers creep model

The Burgers creep model is a widely used mechanical model in rock mechanics, which can effectively describe the complex creep behavior of rocks. It is composed of a Maxwell body and a Kelvin body connected in series (Yang *et al.*, 2023), as shown in Fig. 8, and the creep equation is

$$\varepsilon(t) = \frac{\sigma}{E_1} + \frac{\sigma}{\eta_1}t + \frac{\sigma}{E_2}\left(1 - e^{-\frac{E_2}{\eta_2}t}\right), \tag{4.1}$$

where σ and ε are stress and strain; E_1 and η_1 are the elastic modulus and viscosity coefficient of the Maxwell body; and E_2, η_2 are the elastic modulus and viscosity coefficient of the Kelvin body.

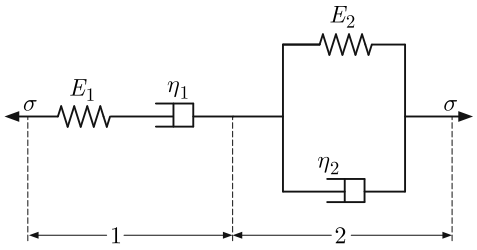

Fig. 8. Burgers model.

However, the Burgers creep model has several inherent limitations (Goodman, 2008; Jaeger & Cook, 2007). Composed of linear elements (a spring-dashpot series in the Maxwell body and a spring-dashpot parallel combination in the Kelvin body), it cannot accurately represent the

accelerating creep stage. The rapid increase in the deformation rate during this stage is a nonlinear behavior that the model's linear components cannot capture, which leads to inaccuracies in predicting full-range creep.

For mudstone, which is highly sensitive to water content changes, the Burgers model fails to account for internal damage evolution (Hudson & Harrison, 1997; Brown, 2004). It does not consider how water-induced micro-structural degradation affects the creep process (Hoek & Brown, 2002; Lemaitre, 1996). Water infiltration softens clay minerals in mudstone, causing water-induced damage and altering its internal structure, but the Burgers model cannot incorporate these changes.

As a result, the model cannot adequately describe the weakening of mudstone's mechanical properties due to water-rock interactions, which limits its accuracy in predicting long-term behavior under different water contents. To address these issues, it is very necessary to define water-induced and creep-induced damage variables.

4.2. Damage variable

Based on damage mechanics theory, the damage variable is defined as the ratio of the defective area of the material to the total effective load-bearing area of the material (Murakami, 1988; Kachanov, 1958). Mathematically, it can be expressed as

$$D = \frac{A_0 - A_w}{A_0} = 1 - \frac{A_w}{A_0}, \tag{4.2}$$

where A_0 is the effective bearing area in the undamaged state, and A_w is the effective bearing area in the damaged state due to the influence of water content.

As shown in Subsection 3.1, the elastic modulus of mudstone follows an exponential decay with the increase in water content. The fitting equation for the elastic modulus E and water content w can be expressed as

$$E_w = E_0 e^{-fw}, \tag{4.3}$$

where E_0 is the elastic modulus of the dry mudstone specimen, and f and w are fitting parameters.

Since the elastic modulus is related to the effective load-bearing area of the material, and the change in elastic modulus with water content reflects the degree of damage to the material caused by water, when the material is damaged by water, the elastic modulus decreases. According to the definition of the damage variable, the water-content-induced damage variable D_w can be expressed as

$$D_w = 1 - \frac{E(w)}{E_0} = 1 - e^{-fw}. \tag{4.4}$$

Meanwhile, mudstone is also affected by creep during long-term service. The creep-induced damage of mudstone is a process of continuous internal structure degradation over time (Rabotnov, 1969; Lemaitre & Chaboche, 1990). In the study of creep-induced damage, we start from the basic concept of damage mechanics. The damage variable is often defined to describe the degree of material degradation.

Let us assume that the rate of damage evolution $\frac{\mathrm{d}D_c}{\mathrm{d}t}$ is proportional to a power-law function of time. Mathematically, it can be written as

$$\frac{\mathrm{d}D_c}{\mathrm{d}t} = kt^{n-1}, \tag{4.5}$$

where k is a proportionality constant and n is a material-specific parameter.

Integrating both sides of the equation with respect to time from 0 to t (Rabotnov, 1969; Lemaitre & Chaboche, 1990):

$$\int_0^{D_c} \mathrm{d}D_c = \int_0^t kt^{n-1}\,\mathrm{d}t, \qquad D_c = k\frac{t^n}{n}\bigg|_0^t, \qquad D_c = \frac{k}{n}t^n. \tag{4.6}$$

Considering the physical meaning that at $t = 0$, $D_c = 0$, and normalizing the damage variable so that D_c ranges from 0 (undamaged) to 1 (completely damaged), after a series of mathematical treatments and in combination with a large number of experimental studies and theoretical analyses, we can assume that $D_c = 1 - e^{-t^n}$. This parameter n can be determined by conducting a series of creep tests on mudstone specimens, and then fitting the experimental data of the creep-induced damage degree changing with time, so as to accurately reflect the development law of creep-induced damage of mudstone.

Considering both the damage caused by water content and creep, the total damage variable, D of mudstone can be expressed as a combination of the two damage variables (Krajcinovic, 1996):

$$D = 1 - (1 - D_w)(1 - D_c). \tag{4.7}$$

This formula comprehensively considers the coupling effect of water-content-induced damage and creep-induced damage on mudstone. As the water content w increases, the value of e^{-fw} decreases, and D_w increases, indicating that the degree of damage to the mudstone caused by water becomes more severe. And as the creep time t increases, the value of e^{-t^n} decreases, and D_c increases, reflecting the continuous development of creep-induced damage. This comprehensive damage variable equation provides a more accurate and comprehensive basis for further establishing the creep damage constitutive model considering the influence of water content, which can more accurately describe the mechanical behavior of mudstone under different water content and creep–time conditions.

4.3. Creep damage constitutive model

To establish a creep damage constitutive model that takes into account the influence of water content and creep behavior (Krajcinovic, 1996), we combine the Burgers model with the damage variable. Figure 9 illustrates the proposed creep damage model, which integrates the Burgers model with both water-induced and creep-induced damage variables. This integration allows a more comprehensive representation of the complex interactions that govern the long-term behavior of mudstone.

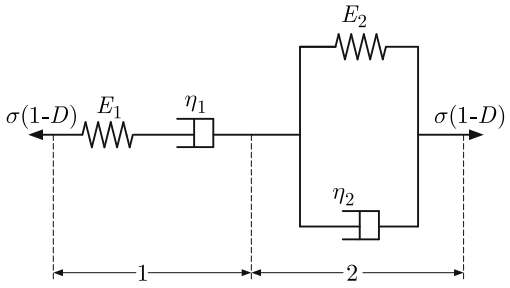

Fig. 9. Illustration of the creep damage model.

The total strain of the mudstone considering damage can be expressed as

$$\varepsilon(t) = (1 - D)\frac{\sigma}{E_1} + \frac{\sigma}{\eta_1}t + \frac{\sigma}{E_2}\left(1 - e^{-\frac{E_2}{\eta_2}t}\right). \tag{4.8}$$

4.4. Model verification

To verify the accuracy and reliability of the established creep damage model, we compared the model's calculated results with our experimental data. Due to the limited number of specimens tested under creep conditions, we selected all creep test specimens for model calibration and validation. This approach maximizes the use of available data to assess the model's performance. The fitting process employed the Quasi-Newton (BFGS) method. We selected several representative specimens with different water contents from the experiment and input the corresponding stress levels and material parameters into the model. Then, we calculated the creep strain at different times using the model and compared it with the measured creep strain from the experiment. The results of this comparison are visually presented in Fig. 10.

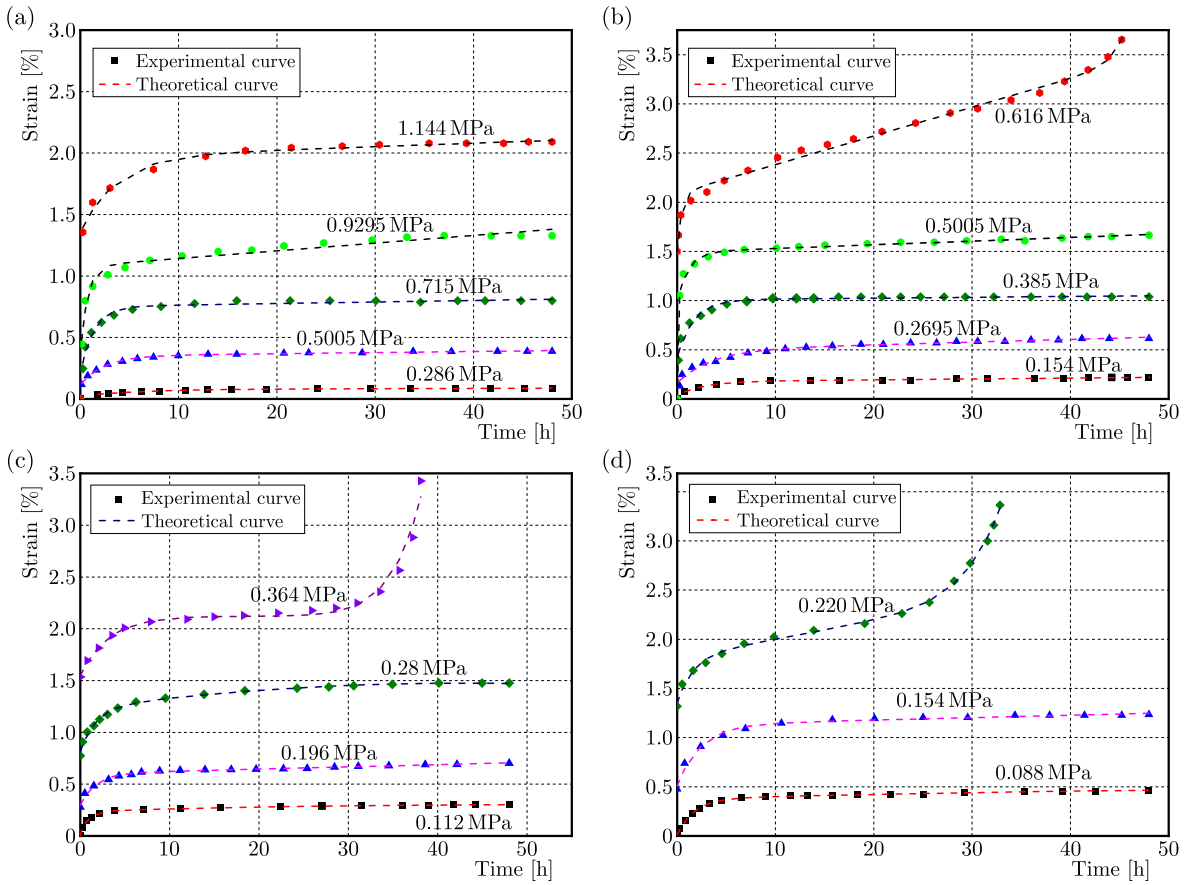

Fig. 10. Comparison between experimental creep curves and creep damage model curves: (a) water content of 0.00 %; (b) water content of 0.26 %; (c) water content of 0.72 %; (d) water content of 1.64 %.

It can be seen from Fig. 10 that the experimental creep curves and the curves simulated by the proposed creep damage model for specimens with different water contents demonstrate a reasonable degree of agreement. The experimental data points are plotted as discrete markers, while the model-predicted curves are shown as continuous lines. Specifically, during the primary creep stage, the model-predicted curves accurately capture the decreasing trend of the creep rate, aligning with the experimental observations. During the secondary creep stage, the predicted creep rates are close to the measured values, indicating that the model can effectively represent the relatively stable creep behavior. In the tertiary creep stage, although minor discrepancies exist between the model-predicted and experimental curves, the model still reflects the overall trend of accelerating creep deformation. The parameters used for the model fitting are listed in Table 3. These parameters were obtained through the fitting process and are specific to each specimen's water content and stress level. This overall agreement between model predictions and

Table 3. Parameters of the creep damage constitutive model.

Water content [%]	Stress [MPa]	E_1 [MPa]	η_1 [MPa · h]	E_2 [MPa]	η_2 [MPa · h]	f	n	k	R^2
0	0.29	13.51	1335.45	5.13	28.46	0.00	24.94	0.00	0.99
	0.50	4.41	546.32	2.12	5.09	0.00	19.21	0.00	1.00
	0.72	23.83	345.17	1.03	0.78	0.00	66.20	0.00	0.98
	0.93	2.57	149.10	1.29	0.94	0.00	30.34	0.00	0.97
	1.14	0.85	406.13	1.86	6.59	0.00	72.17	0.00	0.99
0.26	0.15	1.19E-04	10.00	3.03E-05	9.04E-05	40.087	2.454	−3.14E-05	0.99
	0.27	0.92	10.00	0.46	0.87	2.938	1.012	0.01	0.99
	0.39	4.21E-03	0.32	2.97E-03	3.67E-03	20.945	2.045	1.12E-04	0.98
	0.50	6.83E-03	5.25	109.14	2.31	14.704	−0.491	−8.54E-02	1.00
	0.62	1.16E-02	0.58	2.69E-02	1.11E-02	13.810	35.973	−7.24E-60	1.00
0.72	0.11	1.28	17.11	0.31	0.32	0.90	1.17	3.15E-03	1.00
	0.20	1.40	191.45	1.37	2.30	−1.05	20.77	0.00	0.99
	0.28	0.42	20.47	0.84	1.42	−0.28	1.18	3.38E-03	0.99
	0.36	2.87E-05	45482.50	7.74E-05	2.44E-04	12.52	11.50	−4.10E-18	0.98
1.64	0.09	1.92E-03	2.49	6.68E-05	1.19E-04	4.98	1.00	−4.45E-03	1.00
	0.15	0.29	59.85	0.24	0.57	2.58E-02	88.58	0.00	0.99
	0.22	0.16	11.81	0.48	0.74	−2.54E-04	8.85	−1.27E-13	1.00

experimental measurements, despite some simplifications in the model and potential measurement errors in the experimental data, supports the effectiveness of the proposed creep damage model in describing the creep behavior of mudstone under different water content and stress conditions.

5. Conclusion

This study has comprehensively investigated the long-term mechanical properties of mudstone under different water content conditions. The results have demonstrated that water content has a significant impact on the short-term strength, elastic modulus, and creep behavior of mudstone.

In terms of short-term strength, both the uniaxial compressive strength and elastic modulus of mudstone decrease exponentially with increasing water content. The initial uniaxial compressive strength of dry mudstone was 1.43 MPa, and it dropped to 0.44 MPa when the water content reached 1.64 % (saturation), with a reduction of 69.45 %. The elastic modulus decreased from 14.35 MPa for the dry specimen to 4.12 MPa for the saturated specimen, a reduction of 69.23 %. This exponential decline indicates that water has a strong softening effect on mudstone, which should be carefully considered in engineering designs.

Regarding the creep behavior, water content also plays a crucial role. When the water content is low, the creep deformation of mudstone is relatively small, and the specimen shows some elastic-like behavior in the primary creep stage. However, as the water content increases, the creep deformation increases significantly, the primary creep stage becomes shorter, and the specimen enters the secondary creep stage earlier. Higher water content accelerates the creep failure of mudstone, which poses potential risks to the long-term stability of engineering projects.

To better understand the long-term mechanical behavior of mudstone under varying water content conditions, we proposed expressions for water-induced and creep-induced damage variables and developed a creep damage model based on the combination of these damage variables

and the Burgers creep model. This model can effectively describe the complex creep behavior of mudstone. The verification results showed that the model-calculated values were in good agreement with the experimental data, although there were some errors due to model simplification and experimental measurement errors.

Future research could focus on further improving the model by considering more complex factors such as the interaction between water and other chemical substances in mudstone, as well as the influence of different stress states on creep damage. Additionally, more advanced experimental techniques can be employed to accurately measure the micro-structural changes of mudstone during the creep process, which will help to establish more accurate constitutive models and provide more reliable theoretical support for engineering applications.

References

1. Bieniawski, Z.T., & Bernede, M.J. (1979). Suggested methods for determining the uniaxial compressive strength and deformability of rock materials. *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts*, *16*(2), 135–140.
2. Brown, E.T. (2004). *Rock characterization, testing and monitoring: ISRM suggested methods*. ISRM.
3. Chen, F., Sun, X., & Lu, H. (2022). Influence of water content on the mechanical characteristics of mudstone with high smectite content. *Geofluids*, *2022*(1), Article 9855213. https://doi.org/10.1155/2022/9855213
4. Gao, Y., Wei, W., & Jiang, Q. (2023). Effect of water content on mechanical properties and internal microcrack evolution in mudstone. *Arabian Journal for Science and Engineering*, *48*(10), 12775–12791. https://doi.org/10.1007/s13369-022-07569-9
5. Goodman, R.E. (2008). *Introduction to rock mechanics* (2nd ed.). John Wiley & Sons.
6. Hoek, E., & Brown, E.T. (1997). Practical estimates of rock mass strength. *International Journal of Rock Mechanics and Mining Sciences*, *34*(8), 1165–1186. https://doi.org/10.1016/S1365-1609(97)80069-X
7. Hudson, J.A., & Harrison, J.P. (1997). *Engineering rock mechanics: an introduction to the principles*. Elsevier.
8. Jaeger, J.C., Cook, N.G.W., & Zimmerman, R.W. (2007). *Fundamentals of rock mechanics* (4th ed.). Blackwell Publishing.
9. Kachanov, L.M. (1958). Time of the rupture process under creep conditions. *Izvestiia Akademii Nauk SSSR, Otdelenie Teckhnicheskikh Nauk*, *8*, 26–31.
10. Krajcinovic, D. (1996). *Damage mechanics*. Elsevier.
11. Lemaitre, J. (1996). *A course on damage mechanics*. Springer. https://link.springer.com/book/10.1007/978-3-642-18255-6
12. Lemaitre, J., & Chaboche, J.L. (1990). *Mechanics of solid materials*. Cambridge University Press. https://docs.dicatechpoliba.it/filemanager/407/MECHANICS%20OF%20SOLID%20MATERIALS%20-%20cap%201%20e%203.pdf
13. Li, J., Gao, Y., Yang, T., Zhang, P., Deng, W., & Liu, F. (2023). Effect of water on the rock strength and creep behavior of green mudstone. *Geomechanics and Geophysics for Geo-Energy and Geo-Resources*, *9*(1), Article 101. https://doi.org/10.1007/s40948-023-00638-9
14. Liu, C.-D., Cheng, Y., Jiao, Y.-Y., Zhang, G.-H., Zhang, W.-S., Ou, G.-Z., & Tan, F. (2021). Experimental study on the effect of water on mechanical properties of swelling mudstone. *Engineering Geology*, *295*, Article 106448. https://doi.org/10.1016/j.enggeo.2021.106448
15. Liu, J., Zhang, Q., Wang, L., Chen, F., Wang, P., Yan, X., & Guo, L. (2024). A time-dependent expansion model for mudstone submerged in water. *Soil Mechanics and Foundation Engineering*, *61*(1), 20–26. https://doi.org/10.1007/s11204-024-09938-y

16. Ma, C., Zhan, H.-B., Yao, W.-M., & Li, H.-Z. (2018). A new shear rheological model for a soft interlayer with varying water content. *Water Science and Engineering*, *11*(2), 131–138. https://doi.org/10.1016/j.wse.2018.07.003

17. Murakami, S. (1988). Mechanical modeling of material damage. *Journal of Applied Mechanics (ASME Transactions)*, *55*(2), 280–286. https://doi.org/10.1115/1.3173673

18. Ping, S., Wang, F., Wang, D., Li, S., Wang, Y., Yuan, Y., & Feng, G. (2024). Mechanical damage induced by the water–rock reactions of gypsum-bearing mudstone. *Rock Mechanics and Rock Engineering*, *57*(8), 6377–6394. https://doi.org/10.1007/s00603-024-03855-0

19. Qi, S., Zhang, H., Bian, H., Wang, J., Xu, S., & Wu, B. (2024). Damage characteristics and degradation mechanism of silty mudstone under wet–dry cycling. *Geotechnical and Geological Engineering*, *42*(7), 6095–6112. https://doi.org/10.1007/s10706-024-02877-3

20. Rabotnov, Yu.N. (1969). *Creep problems in structural members*. North-Holland Publishing Company.

21. Sawatsubashi, M., Kiyota, T., & Katagiri, T. (2021). Effect of initial water content and shear stress on immersion-induced creep deformation and strength characteristics of gravelly mudstone. *Soils and Foundations*, *61*(5), 1223–1234. https://doi.org/10.1016/j.sandf.2021.06.015

22. Schimmell, M.T.W., Hangx, S.J.T., & Spiers, C.J. (2022). Effect of pore fluid chemistry on uniaxial compaction creep of Bentheim sandstone and implications for reservoir injection operations. *Geomechanics for Energy and the Environment*, *29*, Article 100272. https://doi.org/10.1016/j.gete.2021.100272

23. Shao, Z., Song, Y., Zheng, J., Shen, F., Liu, C., & Yang, J. (2024). Damage degradation mechanism and macro-meso structural response of mudstone after water wetting. *Journal of Mountain Science*, *21*(8), 2825–2843. https://doi.org/10.1007/s11629-023-8580-x

24. Wang, J.-G., Sun, Q.-L., Liang, B., Yang, P.-J., & Yu, Q.-R. (2020). Mudstone creep experiment and nonlinear damage model study under cyclic disturbance load. *Scientific Reports*, *10*, Article 9305. https://doi.org/10.1038/s41598-020-66245-w

25. Wang, Y., Cong, L., Yin, X., Yang, X., Zhang, B., & Xiong, W. (2021). Creep behaviour of saturated purple mudstone under triaxial compression. *Engineering Geology*, *288*, Article 106159. https://doi.org/10.1016/j.enggeo.2021.106159

26. Yang, S.-Q., Tian, W.-L., Jing, H.-W., Huang, Y.-H., Yang, X.-X., & Meng, B. (2019). Deformation and damage failure behavior of mudstone specimens under single-stage and multi-stage triaxial compression. *Rock Mechanics and Rock Engineering*, *52*(3), 673–689. https://doi.org/10.1007/s00603-018-1622-y

27. Yang, Y.-L., Zhang, T., Liu, S.-Y., & Luo, J.-H. (2022). Mechanical properties and deterioration mechanism of remolded carbonaceous mudstone exposed to wetting–drying cycles. *Rock Mechanics and Rock Engineering*, *55*(6), 3769–3780. https://doi.org/10.1007/s00603-022-02833-8

28. Yang, Y.J., Huang, G., Zhang, Y.Q., & Yuan, L. (2023). An improved Burgers creep model of coal based on fractional-order. *Frontiers in Earth Science*, *11*, Article 1277147. https://doi.org/10.3389/feart.2023.1277147

29. Yu, X.-W., Fu, H.-Y., Zeng, L., Liu, J., & Qiu, X.-Y. (2024). Damage constitutive model for soft rocks and its experimental verification on silty mudstone considering cyclic rock-water interactions. *Bulletin of Engineering Geology and the Environment*, *83*(6), Article 254. https://doi.org/10.1007/s10064-024-03740-8

30. Zheng, J., Song, Y., Shen, F., Shao, Z., Liu, C., & Yang, J. (2024). Study on mechanical properties of water-immersed mudstone based on nanoindentation tests. *Mining, Metallurgy & Exploration*, *41*(4), 2031–2046. https://doi.org/10.1007/s42461-024-01027-w

31. Zou, J., Li, G., Li, Z., Zhang, Y., Liu, H., & Wang, Y. (2024). Experimental study on the mechanical characteristics of weakly cemented mudstone under different loading rates. *Scientific Reports*, *14*, Article 15364. https://doi.org/10.1038/s41598-024-65024-1

Manuscript received March 20, 2025; accepted for publication June 24, 2025; published online November 8, 2025.

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 903–913, Warsaw 2025
https://doi.org/10.15632/jtam-pl/208833

TURBULENT COHERENT STRUCTURES IN THERMAL VORTEX RINGS

Paweł JĘDREJKO

Faculty of Physics, University of Warsaw, Warsaw, Poland
p.jedrejko@uw.edu.pl

The study concerns self-similar structures that emerge during the process of the thermal vortex ring formation. A qualitative explanation of their origin is provided based on the repetitive Kelvin–Helmholtz instability in multiple scales. This phenomenon is found to invert the turbulent energy cascade near the buoyancy interface. To quantify the associated mixing, the fractal dimension of the interface is also computed.

Keywords: thermal; vortex; ring; coherent; inverse cascade.

1. Introduction

Thermal vortex rings are an important feature of atmospheric convection. They rise from buoyancy anomalies, i.e., regions of an increased temperature, transporting energy and moisture upwards. In the final stage, they lead to the formation of cumulus clouds (Yano, 2023).

Due to the very wide range of scales in the atmosphere, thermals are usually left unresolved in numerical weather prediction. However, they are used as conceptual building blocks of subgrid-scale, convective phenomena which have to be modeled. For that reason, features of thermals' dynamics are of high interest and remain an active field of study (Morrison *et al.*, 2023; Yano & Morrison, 2024).

A particularly significant aspect of the dynamics of an isolated thermal is its entrainment rate (Morrison *et al.*, 2023). This problem is directly connected to the features of near-interface turbulence. Being affected by the updraft and the ring formation, turbulence there is hardly homogenous, isotropic, and statistically steady. A promising approach is to focus on its persisting, case-dependent features and symptoms of self-organization. These could be understood as effects of underlying coherent structures whose dynamics locally dominate the flow.

In this article, we study the early stages of the evolution of the vortex ring. The main focus is on understanding the formation of coherent structures which emerge during the ring formation. The setup of the problem is the same as described in (Jędrejko *et al.*, 2024). However, while (Jędrejko *et al.*, 2024) focuses on the methodology and numerics, the work presented below is devoted to the interpretation of physical phenomena.

The main outcome of the article is a phenomenological explanation of a local inversion of the energy cascade in the proximity of a convective structure. This goal justifies the methodology chosen and makes a novel contribution to the studies of atmospheric turbulence.

Section 2 briefly presents the problem and crucial assumptions to make the article comprehensive. Next, Section 3 shows an outline of the ring evolution to provide a physical context for the study of coherent structures. Section 4 justifies some useful simplifications that allow the dynamics to be conceptually understood. Further sections describe the coherent structures, the associated energy transfer, and mixing processes. The latter is done by determining the fractal dimension of the anomaly's interface.

2. Problem statement

The problem under consideration is the evolution of an axisymmetric, buoyancy anomaly (Fig. 1). The focus is on the early stages of the process, which justifies the assumption of azimuthal symmetry, according to Yano and Morrison (2024). The anomaly consists of a region of uniform, increased temperature T_0, which is related to buoyancy by the Boussinesq approximation:

$$\mathbf{b} = -\mathbf{g}\beta(T - T_\infty), \tag{2.1}$$

where β is the thermal expansion coefficient and T_∞ is the reference ambient temperature. The change in buoyancy is assumed to be discontinuous and its shape is initialized as a sphere using cylindrical coordinates $\{\rho, \phi, z\}$:

$$\mathbf{b}(\mathbf{r}, t=0) = \begin{cases} b_0\widehat{z}, & |\mathbf{r}| \le R, \\ \mathbf{0}, & |\mathbf{r}| > R, \end{cases} \tag{2.2}$$

with $b_0 = g\alpha(T_0 - T_\infty)$. The system starts to evolve from rest, i.e., $\mathbf{u}(\mathbf{r}, t=0) = \mathbf{0}$.

Typical scales of atmospheric thermals can be estimated from (Sherwood *et al.*, 2013), which reports $R \approx 10^3$ [m] and $b_0 \approx 10^{-2}$ [m/s^2]. Together with the air's thermal (α) and momentum (ν) diffusivities $\approx 10^{-5}$ [m/s^2], this results in huge Reynolds and Peclet numbers:

$$\mathrm{Re} = \frac{\sqrt{R\,b_0}\,R}{\nu} \approx 10^{10}, \qquad \mathrm{Pe} = \frac{\sqrt{R\,b_0}\,R}{\alpha} \approx 10^{10} \tag{2.3}$$

(Morrison *et al.* (2023) refers to $\mathrm{Re} \approx 10^9$). For that reason, all diffusive processes are neglected. This assumption implies that the buoyancy distribution remains discontinuous and the sharp interface bounding the anomaly can be tracked in Lagrangian fashion:

$$\mathbf{r}(\xi, t) = \begin{bmatrix} \rho(\xi, t) \\ z(\xi, t) \end{bmatrix}, \qquad \xi \in [-\pi/2, \pi/2) \tag{2.4}$$

with the initial condition (Fig. 1a) of

$$\mathbf{r}(\xi, t=0) = R \begin{bmatrix} \sin(\xi) \\ \cos(\xi) \end{bmatrix}. \tag{2.5}$$

Note that the shape of the interface (Eq. (2.4)) does not depend on ϕ due to the symmetry assumed.

The evolution of buoyancy distribution is governed by a simple advection equation:

$$\frac{D\mathbf{b}}{Dt} = 0. \tag{2.6}$$

Using the vorticity equation:

$$\frac{D\omega}{Dt} = (\omega \cdot \nabla)\mathbf{u} + \nabla \times \mathbf{b}, \tag{2.7}$$

it can easily be noted that the anomaly's interface coincides with a vortex sheet. This is because the source term $\nabla \times \mathbf{b}$ gives $\mathbf{0}$ in regions of uniform $\mathbf{b}$, and singularity at the discontinuity. By introducing the vortex sheet strength γ:

$$\gamma(\xi, t)\,\mathrm{d}\xi = \omega\,\mathrm{d}r\,\mathrm{d}z, \tag{2.8}$$

the vorticity equation is reduced to:

$$\frac{\mathrm{d}\gamma}{\mathrm{d}t} = b_0 \frac{\partial z}{\partial \xi}. \tag{2.9}$$

The system is further solved numerically, as described in detail in (Jędrejko *et al.*, 2024), by discretizing the vortex sheet with a set of nodes and segments connecting the nodes. The time integration is done by an adaptive 4th-order Runge–Kutta scheme. The spatial derivative in Eq. (2.9) is computed with the 2nd-order central difference, and the integral (Eq. (2.10)) with the trapezoidal rule. The latter two allow more flexibility in the adaptive discretization of the sheet than higher-order schemes. Such a procedure is necessary to keep the resolution fine, by splitting segments, which got too long.

An important part of the method is also the regularization of the Biot–Savart kernel, introduced by Krasny (1986) and Nitsche and Krasny (1994):

$$\mathbf{u}_0 = \int_0^{2\pi} \int_{-\pi/2}^{\pi/2} \frac{\gamma \widehat{\phi} \times (\mathbf{r}_0 - \mathbf{r})}{(|\mathbf{r}_0 - \mathbf{r}|^2 + \delta^2)^{3/2}} \rho \, \mathrm{d}\xi \, \mathrm{d}\phi. \tag{2.10}$$

It can conceptually be understood as assigning some finite thickness δ to the vortex sheet. As a result, dumping is applied to the highest wavenumbers of the velocity field induced. The main consequence is the bound on the smallest scales present in the flow, especially the smallest wavelengths of the vortex sheet instabilities. The qualitative evolution of the process remains the same, although the rising speed of the thermal is affected. However, this influence is weak ($<5\,\%$ for $\delta \in [0.004,\ 0.016]$) and separate from the interscale energy transfer. As can be deduced from Eq. (2.10), the impact of δ is mainly local.

Alternative approaches to regularization can also be found in the literature (comparison might be found in (Sohn, 2014)), although no discrepancies, significant for this study, are reported. The advantage of the regularization type chosen is its simplicity. This allows us to take the azimuthal integral analytically and use some algorithmic optimizations (like Dynnikova (2009)), described in detail in (Jędrejko *et al.*, 2024).

3. Outline of the system evolution

In the initial stage, the anomaly experiences rapid collapse at the bottom, which transforms the initial sphere to the final vortex ring. Meanwhile, the vortex sheet at the sides is dominated by a series of coherent vortices, which are the main focus of this article. As time passes, the vortices get larger and fill the "interior" of the anomaly with the vortex sheet. This is done by intensive stretching and folding. Figure 1 presents successive stages of the ring's evolution. The subfigure (a) captures the initial condition, (b) the beginning of the collapse with coherent vortices on the side, and (c) the beginning of the ring's closure and space-filling interface.

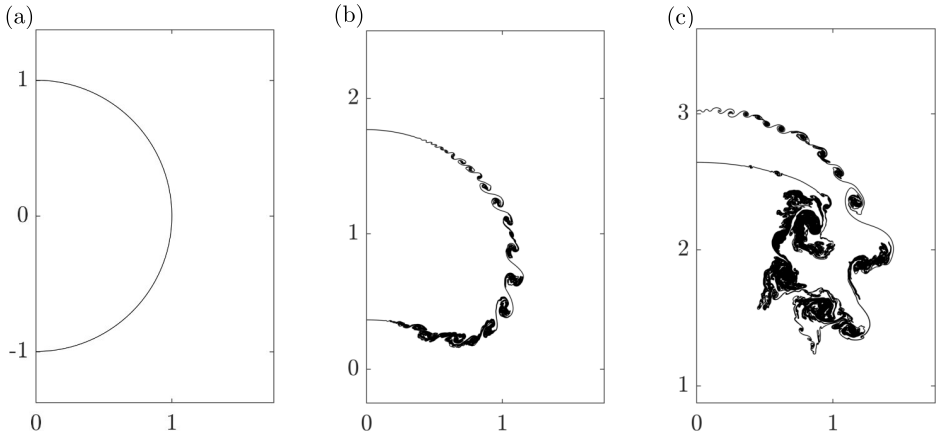

Fig. 1. Evolution of the interfacial vortex sheet in selected time steps: (a) $t = 0$; (b) $t \approx 1.5$; (c) $t \approx 3$. Obtained with $\delta = 0.008$.

4. An analogy to Kelvin–Helmholtz instability

The vortex sheet strength is initially amplified at the sides by the buoyancy (2.9), which launches two concurrent processes (Fig. 1b). The first is the anomaly collapse at the bottom, which ultimately turns it into a vortex ring. The second takes place at the sides and leads to the formation of coherent vortices exhibiting some self-similarity features.

In this section, we will argue that the latter can be qualitatively understood in analogy to the classical Kelvin–Helmholtz (K-H) instability. By that, we mean the case of the plane, periodic vortex sheet, with constant strength γ and finite thickness δ. Such a sheet experiences a roll-up when perturbed (Vallis, 2006 [chapter 6.2.4]; Krasny, 1986), giving rise to the characteristic "cat-eye" vortices.

The circumstances of vortex formation considered in this article differ from the classical K-H by a few features. The vortex sheet is curved, its strength is dynamically changed by buoyancy, and the simultaneous collapse exerts stretching. However, we will argue in favor of a scale separation, which leaves the coherent vortices relatively unaffected by these aspects.

The characteristic length of the initial K-H vortices is δ from Eq. (2.10), which is shown in (Krasny, 1986). That reference discusses perturbations in the form of:

$$x = Xe^{\sigma t + ik\gamma\xi}, \qquad y = Ye^{\sigma t + ik\gamma\xi}, \tag{4.1}$$

where $\{x, y\}$ are Cartesian coordinates describing the shape of the vortex sheet, X, Y are constant, initial amplitudes and $k = 2\pi/\lambda$ is the wavenumber. The analysis leads to the relation:

$$\sigma^2 = \frac{k(1 - e^{-k\cosh^{-1}(1+\delta^2)})e^{-k\cosh^{-1}(1+\delta^2)}}{4\delta(2+\delta^2)^{1/2}}, \tag{4.2}$$

which we use to numerically obtain the fastest-growing wavelength Λ as a function of δ. The resulting relation seems to be linear as shown in Fig. 2 and by regression found to be:

$$\Lambda(\delta) = \alpha\delta, \qquad \alpha = 6.1445. \tag{4.3}$$

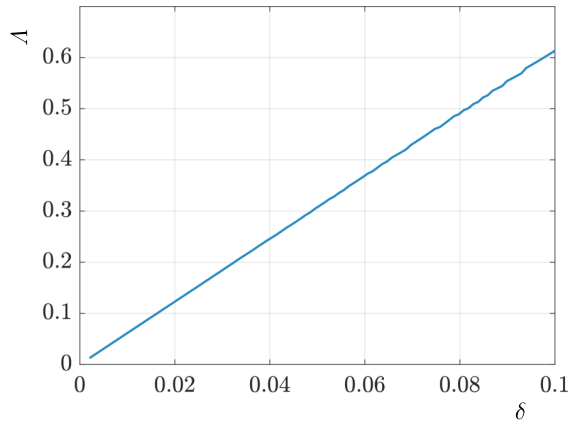

Fig. 2. Fastest growing wavelength as a function of δ, according to Eq. (4.2).

The values of δ considered are much smaller than the anomaly radius:

$$\frac{\delta}{R} \ll 1, \tag{4.4}$$

thus we assume that the curvature of the vortex sheet does not affect the formation of K-H vortices much.

The characteristic time of the anomaly bulk evolution is independent of δ, as shown in (Jędrejko *et al.*, 2024) and can be expressed by

$$t_b = \sqrt{\frac{R}{b_0}}. \tag{4.5}$$

The characteristic time of local γ amplification (Eq. (2.9)) depends on the local shape. Thus, it is t_b before the instability and

$$t_\gamma = \sqrt{\frac{\delta}{b_0}} \tag{4.6}$$

afterward. The cat-eye eddy turn-over time is

$$t_\delta = \frac{\delta^2}{\gamma} = \frac{\delta^2}{b_0 R\, t_b} = \frac{\delta^2}{\sqrt{b_0 R^3}}, \tag{4.7}$$

where we used Eq. (2.9) to determine the accumulation of γ till the emergence of K-H eddies. This happens in time t_b in the condition of local shape characterized by R.

As the first outcome:

$$\frac{t_\delta}{t_\gamma} = \left(\frac{\delta}{R}\right)^{3/2} \ll 1, \tag{4.8}$$

so it is expected that $\partial\gamma/\partial t$ is of secondary importance for the evolution of K-H vortices at the sides of the anomaly.

This result was also checked numerically by running a separate simulation with γ fixed in time:

$$\gamma_0 = b_0 \cos(\xi), \tag{4.9}$$

which is an initial tendency of $\gamma(\xi, t)$, deduced from Eqs. (2.9) and (2.5). The comparison is presented in Fig. 3 for the snapshots, where the anomaly center is at the same height. This happens for $t \approx 0.56$ for the constant γ_0 and $t \approx 0.89$ for dynamic γ. The time shift is due to the fact that in the latter case, γ has to be amplified in time to reach the value of γ_0. This happens mostly "in place", because the initial γ is too weak to significantly change the state of the system. However, the tiny progress of the anomaly's collapse in the initial period results in a small difference in its thickness along the vertical axis (Fig. 3a). Despite these differences in the bulk evolution and the resulting "rigid-body" translation of the K-H vortices, their shapes are

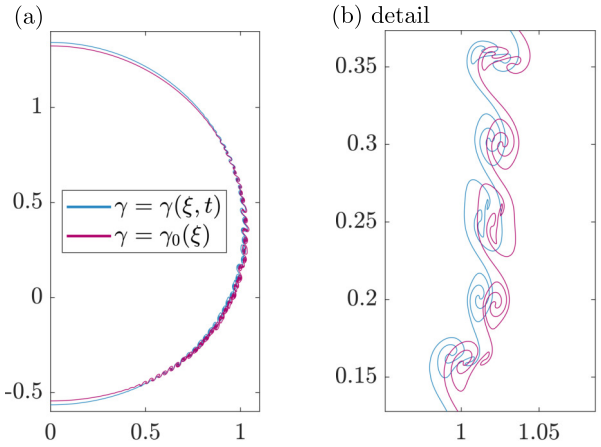

Fig. 3. Comparison of the case with γ evolving according to Eq. (2.9) and the case with γ fixed at the initial tendency. Anomalies centered at the same point.

de facto indistinguishable (Fig. 3b). Therefore, the time dependence of γ is insignificant for the evolution of K-H vortices.

The second outcome from the dimensional analysis is

$$\frac{t_\delta}{t_b} = \left(\frac{\delta}{R}\right)^2 \ll 1, \tag{4.10}$$

so the collapse and the initial cat-eye vortices have well-separated time scales.

Knowing already that the sheet curvature (Eq. (4.4)) and $\partial\gamma/\partial t$ (Eq. (4.8)) are negligible for K-H, an a posteriori argument for the collapse and K-H separation is the accuracy of Eq. (4.3). For $\delta = 0.008$ it predicts $\Lambda = 0.04916$ and, as shown in Fig. 5, despite the collapse we get $\Lambda = 0.04908$.

In summary, Eqs. (4.4), (4.8), (4.10) justify the reasoning based on the classical K-H instability in understanding the coherent structures at the sides of the anomaly.

5. The concept of hierarchical Kelvin–Helmholtz instability

The most interesting feature of the coherent structures on the sides of the anomaly is their self-similarity. We explain it by referring to the idea of hierarchical instability.

First, K-H instability occurs, and the vortex sheet gets covered with a layer of cat-eye vortices. Their size is determined by the sheet thickness δ as given by Eq. (4.3). Such a layer effectively starts to behave like a new, thicker vortex sheet. Because it is built of smaller structures, its effective strength is initially perturbed. This leads to the new K-H instability in higher wavelengths due to higher effective thickness. The process repeats, with each new generation of vortices approximately doubling the characteristic wavelength of the previous one. This proceeds till the value of effective δ breaks the condition (4.4), then (4.8) and (4.10), which couples the dynamics of the structures with other processes in the system.

This interpretation is justified by running the case with twice the higher value of δ. The resulting vortices are very similar to the second generation of vortices from the case with lower δ, see Fig. 5. Both systems further evolve analogically, doubling the characteristic size of the structures in an iterative manner. This phenomenon naturally raises a question about the local inverse energy cascade.

6. Energy transfer by the hierarchical K-H

The investigation of the energy transfer associated with the structures described in the previous section is troublesome. The two-dimensional Fourier transform would have to be bounded to a finite, non-periodic domain. It is also highly affected by the updraft in the center of the anomaly. To analyze the energy of the interfacial structures exclusively, we turn to different methodology based on an FFT along the contour.

6.1. Generation renewal

The vortex sheet is parametrized with its initial length, as in Eq. (2.4). Note that although the formation of a single cat-eye vortex stretches the sheet very intensively, it takes place in a fixed range of ξ.

As long as the overall shape of the anomaly was not affected much by the collapse (say $t < 1.5$, compare with Fig. 1), this range is a good measure of the vortex size. This is because it corresponds to the length of that interface piece at the reference stage, i.e., before the roll-up (Fig. 4).

This approach also holds for the next generations of vortices, as long as the collapse does not proceed too far. For that reason, a Fourier transform of functions of ξ provides an insight

Fig. 4. Conceptual drawing of a material piece of the interface before (light) and after (dark) the roll-up. Its initial length $(\xi_2 - \xi_1)$ is a good measure of the eddy size.

into eddy scale distribution. This is especially convenient because an interface, as a closed loop, is periodic. Fortunately, scales of the anomaly collapse and K-H (in the initial stage) are well separated. We use a threshold of $\lambda = 0.5$ and interpret the most energetic wavelength below it as a characteristic scale of the K-H. Figure 5 presents a dominant K-H scale as a function of time for three values of δ.

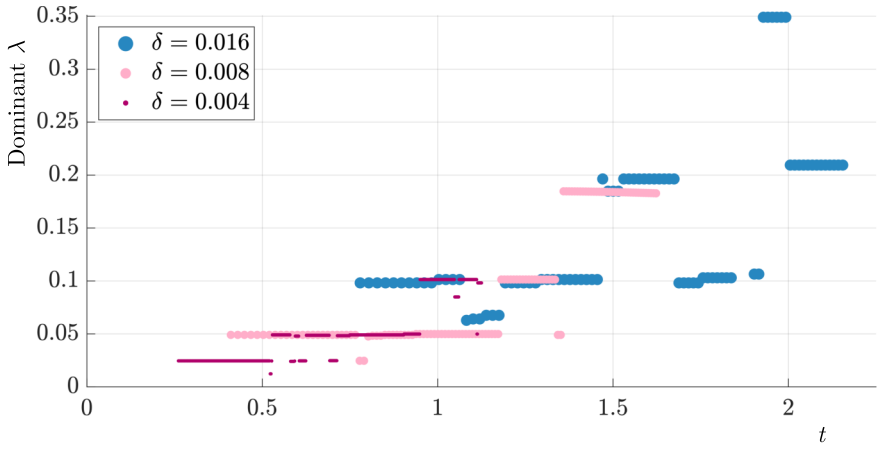

Fig. 5. Most energetic wavelengths (among shorter than 0.5) in time show distinct generations of vortices.

The instability seems to develop later for higher δ, but this can be caused by the details of discretization applied to each case. What is worth noting is a clear generation renewal with size doubling, leading to the shift of energy towards large wavelengths. Moreover, the second generation for δ closely matches the scale of the first generation for 2δ. In further times $t > 1.5$, the size doubling is less exact. This might be caused by the collapse, which makes the initial condition no longer a good reference point, or by breaking conditions (4.4), (4.8), (4.10). At some stages, the competition between current and previous generations is close, leading to a temporal jump-back of the dominant λ. This indicates that a new generation is built on top of the previous one rather than instead of it. Figure 5 is also in good agreement with Eq. (4.3), which is presented in Table 1.

Table 1. Wavelengths of subsequent generations (Fig. 5) compared with predictions of Eq. (4.3).

δ	Generation 1 (Eq. (4.3))	Generation 1	Generation 2	Generation 3
0.004	0.02458	0.02454	0.04909	0.10134
0.008	0.04915	0.04908	0.10139	0.18467
0.016	0.09831	0.09820	0.19635	0.34907

6.2. Spectrum along the smoothed contour

An alternative approach for the energy transfer analysis is to sample the velocity along a smoothed contour and then compute its Fourier transform. This is more computationally demanding but is not limited to the early stages of the system evolution.

A separate simulation with high $\delta = 0.1$ is used to obtain a smooth contour. Because a significant range of small scales is dumped, the rising and collapse speeds are affected. For that reason, a case with $\delta = 0.1$ at time t does not fit the case $\delta = 0.008$ at t very well. However, if the thickness of the ring (along the z-axis) is matched and the height is adjusted, two contours match closely (example in Fig. 6).

Fig. 6. Smoothed contour ($\delta = 0.1$) in black, in front of the $\delta = 0.008$ interface (gray). Contours associated with the thickness at z-axis. Shifting along z applied to match the heights.

Spectra of low-δ sheet's kinetic energy, computed along high-δ contours, are presented in Fig. 7. The system accumulates energy over time. Therefore, to compare its various stages, plots were normalized with K-H peak energy. Figure 8 includes examples of the flow structures as a reference. The results from Fig. 7 are in good agreement with outcomes from the previous subsection, shown in Fig. 5. They both indicate a transfer of energy towards larger scales. The characteristic wavelengths are also consistent with what can be noted with the naked eye (Fig. 8). This applies to both their size and complexity, which is related to the width of a given peak from Fig. 7.

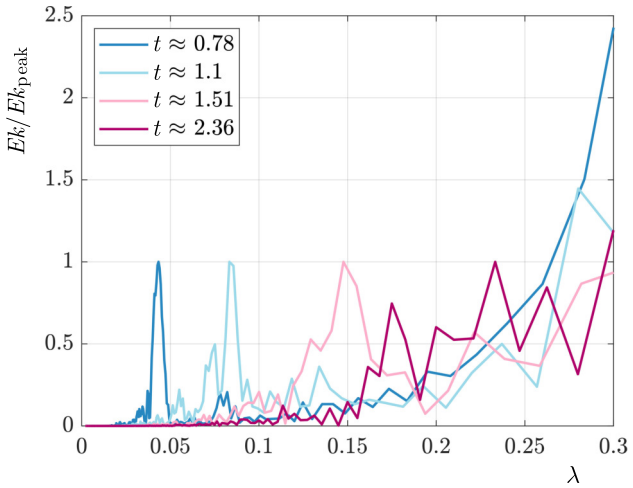

Fig. 7. Normalized energy spectra for selected timesteps. Results obtained with $\delta = 0.008$. Presented wavelength range associated with K-H instability.

7. Fractal dimension of the hierarchical K-H

The intense stretching and folding associated with the hierarchical K-H tightly fills the space with the vortex sheet. The phenomenon is similar in nature to the classical Smale's horseshoe map (Shub, 2005). To measure the intensity of mixing associated with this process, we determine the time evolution of the interface fractal dimension. This is done with the box-counting method

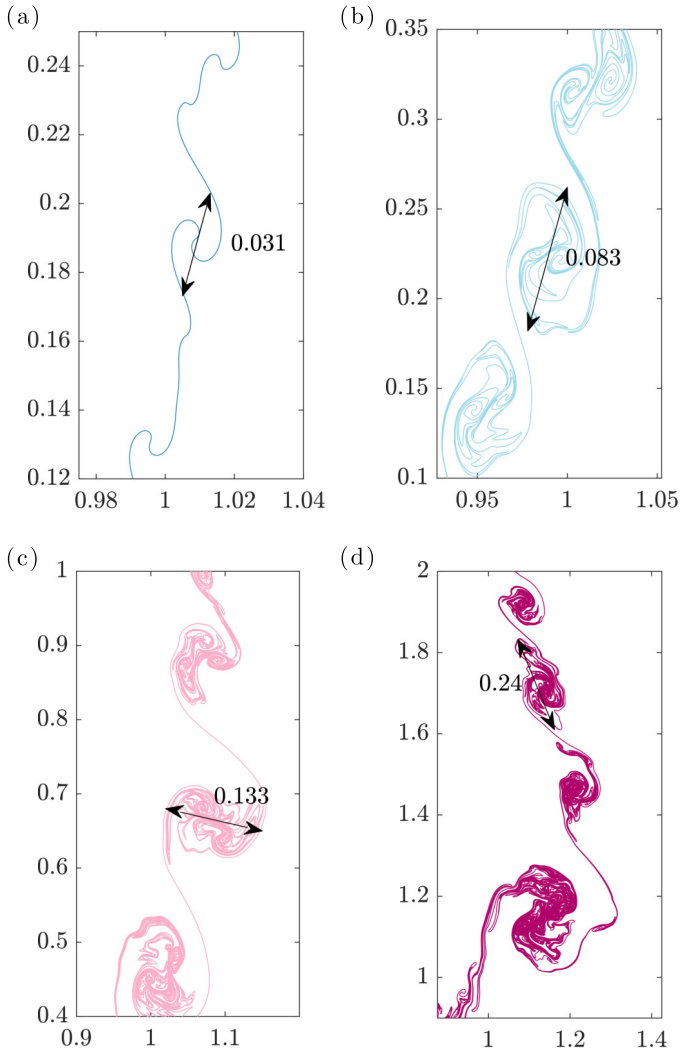

Fig. 8. Approximated characteristic scales of proceeding generations of K-H vortices, $\delta = 0.008$: (a) $t \approx 0.78$; (b) $t \approx 1.10$; (c) $t \approx 1.51$; (d) $t \approx 2.36$. Colors matched with Fig. 7.

(Liebovitch & Toth, 1989). For that purpose, the domain is covered with a uniform grid of spacing (i.e., box side) d. Then, boxes crossed by the vortex sheet are counted, giving a total number of $n(d)$. The process is repeated for a range of box sizes d, and the fractal dimension is evaluated using Minkowski–Bouligand definition (Bishop & Peres, 2017):

$$D_{\text{box}} = \lim_{d \to 0} \frac{\log(n)}{\log(1/d)}, \tag{7.1}$$

which implies:

$$n(d) \approx C d^{-D_{\text{box}}}. \tag{7.2}$$

The scaling of Eq. (7.2) for example timesteps is presented in Fig. 9a. The time evolution of the fractal dimension is shown in Fig. 9b.

The evaluated fractal dimension experiences rapid growth when the hierarchical K-H starts and converges to about 1.78. This exceeds the range of 1.3–1.66 found in cloud interfaces (Malinowski & Zawadzki, 1993). This could be an artifact of axial symmetry, which is less and less justified in later times. The first notable plateau in the fractal dimension (Fig. 9b) is at the level of 1.6, so inside the typical cloud range.

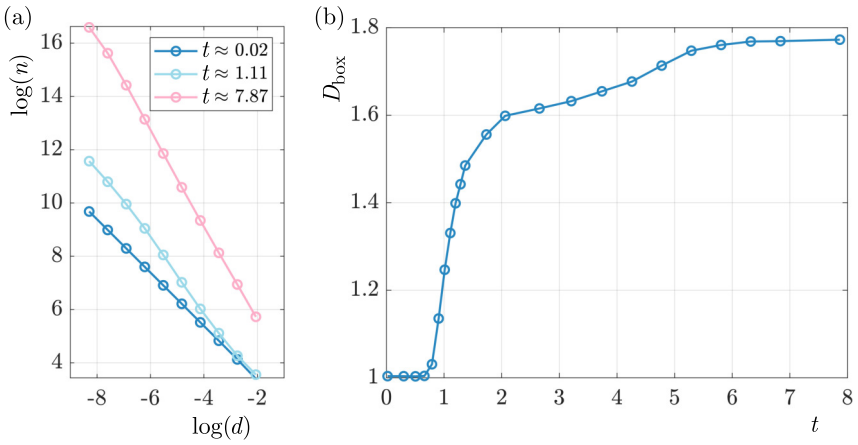

Fig. 9. (a) Number of boxes vs box size; (b) box counting dimension in time.

8. Summary and discussion

In this study, we investigated the coherent structures that emerge in the initial stages of the thermal vortex ring formation. Its self-similar nature was explained in analogy to the K-H instability, which occurs multiple times in increasing wavelengths. A related, subsequent stretching and folding introduce intense mixing. This is manifested as an increase in the interface's fractal dimension, growing up to about 1.78.

This hierarchical K-H instability was also found to locally transfer energy to large scales. Such behavior is characteristic of two-dimensional turbulence (Davidson, 2015, chapter 10) and, for late times, could be an artifact of axial symmetry. However, the time range considered in this paper is definitely within the range of physically justified axial symmetry, according to (Yano & Morrison, 2024). The initial inversion of the turbulent cascade near the interface is, therefore, trustworthy. The important question is how long such an inverse cascade remains active. The problem of its azimuthal stability and interaction with long-term stretching is left for further study.

Acknowledgments

The oral presentation of this work was awarded 2nd place in the Janusz W. Elsner Competition at the 26th Fluid Mechanics Conference in Warsaw (2024). The author gratefully acknowledges the Organizing Committee for recommending this manuscript and covering the publication fee.

References

1. Bishop, C.J., & Peres, Y. (2017). *Fractals in probability and analysis*. Cambridge Studies in Advanced Mathematics, vol. 162. Cambridge University Press.
2. Davidson, P.A. (2015). *Turbulence: An introduction for scientists and engineers* (2nd ed.). Oxford University Press.
3. Dynnikova, G.Ya. (2009). Fast technique for solving the N-body problem in flow simulation by vortex methods. *Computational Mathematics and Mathematical Physics*, *49*(8), 1389–1396. https://doi.org/10.1134/S0965542509080090
4. Jędrejko, P., Yano, J.-I., & Wacławczyk, M. (2024). A Lagrangian approach to thermal vortex rings simulation in high Re and high Pe limit. Under consideration in Theoretical and Computational Fluid Dynamics.
5. Krasny, R. (1986). Desingularization of periodic vortex sheet roll-up. *Journal of Computational Physics*, *65*(2), 292–313. https://doi.org/10.1016/0021-9991(86)90210-X

6. Liebovitch, L.S., & Tibor, T. (1989). A fast algorithm to determine fractal dimensions by box counting. *Physics Letters A*, *141*(8–9), 386–390. https://doi.org/10.1016/0375-9601(89)90854-2

7. Malinowski, Sz., & Zawadzki, I. (1993). On the surface of clouds. *Journal of the Atmospheric Sciences*, *50*(1), 5–13. https://doi.org/10.1175/1520-0469(1993)050%3C0005:OTSOC%3E2.0.CO;2

8. Morrison, H., Jeevanjee, N., Lecoanet, D., & Peters, J.M. (2023). What controls the entrainment rate of dry buoyant thermals with varying initial aspect ratio? *Journal of the Atmospheric Sciences*, *80*(11), 2711–2728. https://doi.org/10.1175/JAS-D-23-0063.1

9. Nitsche, M., & Krasny, R. (1994). A numerical study of vortex ring formation at the edge of a circular tube. *Journal of Fluid Mechanics*, *276*, 139–161. https://doi.org/10.1017/S0022112094002508

10. Sherwood, S.C., Hernández-Deckers, D., Colin, M., & Robinson, F. (2013). Slippery thermals and the cumulus entrainment paradox. *Journal of the Atmospheric Sciences*, *70*(8), 2426–2442. https://doi.org/10.1175/JAS-D-12-0220.1

11. Shub, M. (2005). What is ...a horseshoe?. *Notices of the American Mathematical Society*, *52*(5), 516–517. https://www.ams.org/notices/200505/what-is.pdf

12. Sohn, S.-I. (2014). Two vortex-blob regularization models for vortex sheet motion. *Physics of Fluids*, *26*(4), Article 044105. https://doi.org/10.1063/1.4872027

13. Vallis, G.K. (2017). *Atmospheric and oceanic fluid dynamics. Fundamentals and large-scale circulation* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781107588417

14. Yano, J.-I. (2023). *Geophysical convection dynamics*. Elsevier.

15. Yano, J.-I., & Morrison, H. (2024). Thermal vortex ring: Vortex-dynamics analysis of a high-resolution simulation. *Journal of Fluid Mechanics*, *991*, A18. https://doi.org/10.1017/jfm.2024.485

Manuscript received November 30, 2024; accepted for publication July 31, 2025; published online November 7, 2025.

JOURNAL OF THEORETICAL
AND APPLIED MECHANICS
63, 4, pp. 915–928, Warsaw 2025
https://doi.org/10.15632/jtam-pl/208836

SIMULATED CALCULATION AND APPLICATION OF ANNULAR PRESSURE LOSS FOR DEEP SLIM-HOLE SIDETRACKING HORIZONTAL WELL

Zaiming WANG[1], Yuanyuan SHEN[1], Yi HOU[1], Yanna KAN[1], Weian HUANG[2*], Yuqing ZHU[2]

[1] *Production Technology Research Institute, PetroChina Jidong Oilfield Company, Tangshan, China*
[2] *School of Petroleum Engineering, China University of Petroleum (Huadong), Qingdao, China*
*corresponding author, *masterhuang1997@163.com*

This article analyzes the influence of different displacement (5 L/s–17 L/s (liter per second)) and rotational speed (0 rev/min–120 rev/min) conditions on the annular pressure loss of a slim hole under different eccentricity (0 %–40 %) models through simulation methods and the difference in the annular pressure drop gradient at different drilling tool combinations. Based on numerical simulation, results fitted the multi-factor dimensionless annular pressure drop gradient factor. The accuracy of the fitted factors was verified by calculating the pump pressure of a horizontal wellbore section based on the historical data of the SY-3 well with an error of less than 10 %.

Keywords: slim-hole well; deep sidetracking well; eccentric rotation; annular pressure loss.

1. Introduction

The shale oil geological resources in the Qintong Depression of the Subei Basin are abundant. The SY-3 well has been deployed in order to explore low-cost engineering processes and promote the efficient development of shale oil. This well was drilled through a discontinued old well with sidetrack drilling, using the horizontal drilling technology of a ∅118 mm small wellbore. It is the first deep shale oil small wellbore lateral drilling horizontal well in the work area. Due to factors such as small drilling size and small annular clearance, the annular pressure consumption during the drilling process is high, and the formation is prone to collapse and leakage. Therefore, the pump pressure is limited. Moreover, it has a long horizontal segment and a local upward trajectory, which poses a challenge to wellbore cleaning (Delwiche *et al.*, 1992; Song *et al.*, 2004). There are significant differences in hydraulics between slim-hole drilling and conventional wellbore drilling. Therefore, scholars have conducted theoretical and experimental research on the calculation of eccentric rotation annular pressure drop in small wellbore drilling (McCann *et al.*, 1995; Haciislamoglu & Cartalos, 1994; Cartalos *et al.*, 1996; Hansen *et al.*, 1999; Hemphill & Ravi, 2005; Enfis *et al.*, 2011; Kelessidis *et al.*, 2011; Reed & Pilehvari, 1993; Letelier *et al.*, 2017; Tian *et al.*, 2022; Khatibi *et al.*, 2018; Vieira *et al.*, 2014; Sotoudeh & Frigaard, 2024; Shi & Zhang, 2025; Resell *et al.*, 2025). Various computational models have been established, but all of them have limitations, and all of them have some errors in performing prediction calculations. Experimental studies have demanding requirements on the precision of the instruments and the accuracy of the operation. McCann *et al.* (1995) conducted experimental research on pressure changes in narrow annular spaces, but did not consider the effect of eccentricity. The computational model proposed by Haciishamoglu and Cartalos (1994) and Kelessidis *et al.* (2011) did not take into account the influencing factors of drill string rotation.

Hansen *et al.* (1999) did not analyze the gradient changes in annular pressure drop at sudden diameter changes such as drilling tool joints. Khatibi *et al.* (2018) established a computational model based on experimental research, and did not alter the diameter ratio during the experiment. Sotoudeh and Frigaard (2024) focus on cementing, and the empirical size is not a small wellbore. Resell *et al.* (2025) investigate fluid forces and viscous torque on an inner cylinder that simultaneously rotates about its own axis and orbits within an outer cylinder. However, they did not establish an expression for annular pressure drop. In this regard, the author combines domestic and international research data on small-borehole technology, and further investigates the effects of eccentricity, rotation, displacement, drilling tool combinations, and other factors on the annular pressure consumption of small boreholes through simulation. The fitted multifactorial uncaused annulus pressure drop gradient factor can provide a reference for calculating the annulus pressure consumption and optimizing the hydraulic parameters for drilling small boreholes.

2. Establishment of annular model for slim hole

2.1. Solidworks modeling and fluid domain partitioning

According to the field conditions and drilling design, the commonly used drilling tools combination for drilling in a particular well section is simulated by Solidworks. The simulated pipe string with a length of 25.28 meters is established as ∅118 mm polycrystalline diamond compact bit (PDC) + ∅95 mm measurement while drilling (MWD) + ∅73 mm drill pipe + ∅105 mm joints. The annulus is set to ∅150 mm. The fluid domain is divided by a Boolean operation. The inner pipe column and fluid domain are as shown in Fig. 1.

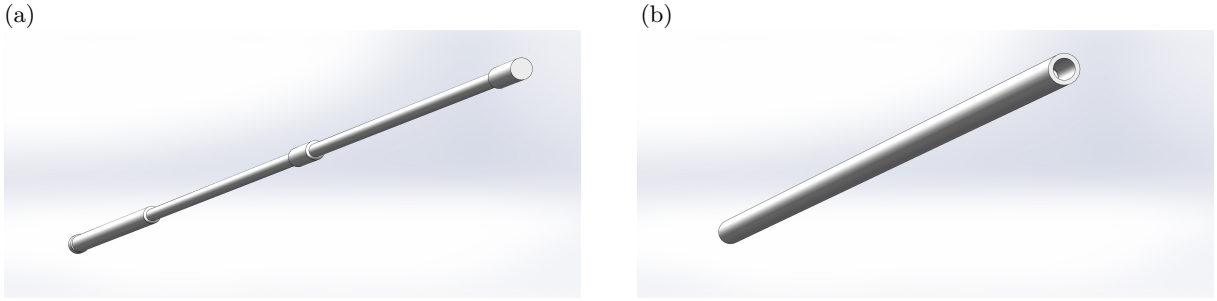

Fig. 1. Inner column and fluid domain: (a) inner pipe; (b) fluid domain.

The eccentricity formula can be expressed as (Tian *et al.*, 2022):

$$E = \frac{\delta}{r_w - r_d}, \tag{2.1}$$

where E is the eccentricity, δ is the eccentricity distance, which is the distance between the drill bit and the two centers of the wellbore in the view along the axis of the wellbore [m], r_w is the radius of the wellbore [m], and r_d is the radius of the drill pipe [m].

The pipe string model is uniformly modeled by Solidworks with five eccentricities: 0 % (eccentricity 0 mm), 10 % (eccentricity 3 mm), 20 % (eccentricity 6 mm), 30 % (eccentricity 9 mm), and 40 % (eccentricity 12 mm), and the eccentricity model is shown in Fig. 2.

Fig. 2. Different eccentricity models: (a) eccentricity 0 %; (b) eccentricity 10 %; (c) eccentricity 20 %; (d) eccentricity 30 %; (e) eccentricity 40 %.

2.2. Fluent meshing and model setup

The model built by Solidworks is imported into Design Modeler, named fluid domain import and export, and structured meshing is performed to accelerate convergence. Taking the 10 % eccentricity model as an example, the meshing results are shown in Fig. 3. A total of 245131 faces were divided with a maximum twist of 0.2174 and a maximum aspect ratio of 16.57, and a total of 546482 control body meshes were divided with a minimum orthogonal mass of 0.3056, a maximum aspect ratio of 5.89, and no isolated meshes. The maximum mesh distortion is 0.6944, the minimum distortion is 0.0029, and the average distortion is 0.0476. The mesh quality is excellent, which meets Fluent's requirements for the mesh quality, and is conducive to the convergence of the calculation.

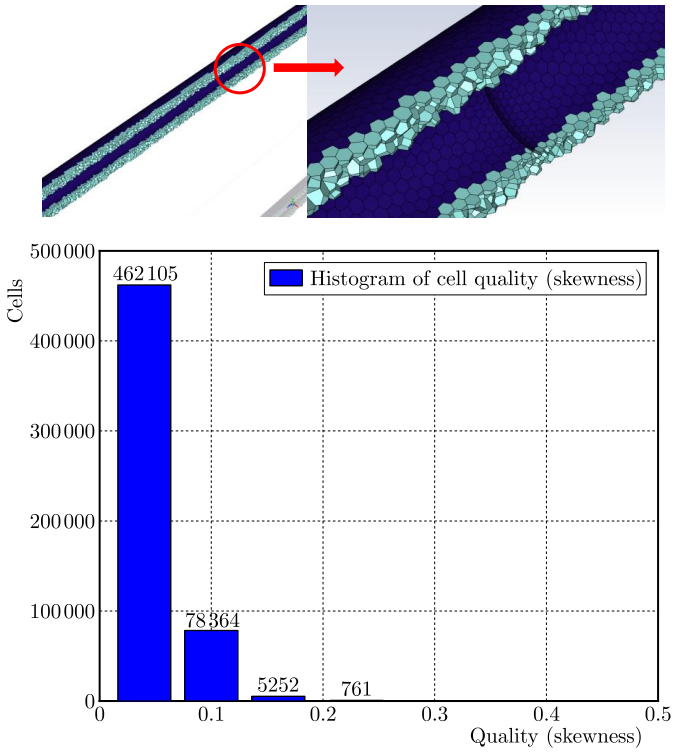

Fig. 3. Physical model and meshing of eccentric annulus with the eccentricity of 10 %.

Let us set the direction of gravity as the y-axis and the direction of the ring-space axis as the z-axis. For different eccentricity models (0 %, 10 %, 20 %, 30 %, 40 %), we set different displacements: 5 L/s, 8 L/s, 11 L/s, 14 L/s, 17 L/s. Numerical simulations were carried out by setting different speeds at different displacements: 0 rev/min, 30 rev/min, 60 rev/min, 90 rev/min, 120 rev/min. Let us calculate the gradient of the pressure drop in the annulus of the whole model at different eccentricities and at different displacements and speeds. The gradients of pressure drop in the annulus at different combinations of drilling tools at different rotational speeds and displacements were calculated for the models with 0 % eccentricity and 10 % eccentricity. For the 10 % model of eccentricity, the gradient of pressure drop in the annulus was calculated with and without a drilling tool joint at different displacements.

The main simulation parameters are set as: drilling fluids set to H-B fluids. According to the performance data of drilling fluid used in the field of SY-3 well, the reference viscosity is 0.058 Pa · s. The yield stress of drilling fluid is 7.15 Pa with a Power Law Index of 0.53. The Consistency Index is 0.48. Drilling fluid is an incompressible fluid with a density set to a constant 1400 kg/m^3. The reference temperature is set at 400 K based on geologic information. Let us set boundaries for the inlet velocity and outlet pressure. The inlet speed is set to 0.139 m/s – 0.556 m/s (5 L/s – 17 L/s) and the rotational speed is 0 rev/min–120 rev/min. The Solver parame-

ters are configured as follows: pressure-velocity coupling using the SIMPLE scheme of predicting the velocity field and then correcting it by the pressure field. To improve the convergence speed and computational accuracy, the spatial discretization format is chosen as the second-order windward format.

3. Analysis of numerical simulation results

3.1. Hydraulic behavior law of eccentric annulus in small wellbore

Eccentric flow fields have asymmetric flow and uneven velocity distribution compared to concentric flow fields. Furthermore, due to the small gap in the small borehole annulus, the drilling fluid is forced to rotate with the drill column due to its viscosity, causing the drilling fluid and drill cuttings spiral flow in the annulus. The flow of drilling fluid in the annulus has been changed to a spiral flow, which is shown in Fig. 4 (Delwiche *et al.*, 1992).

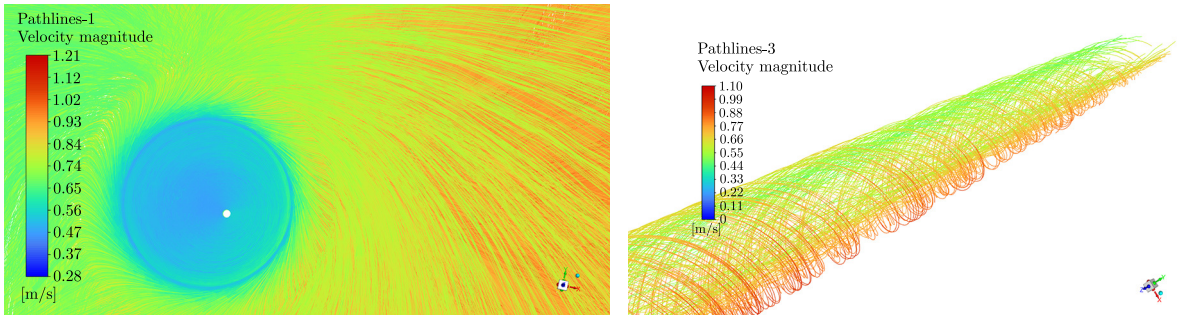

Fig. 4. Simulated slim-hole annular fluid flow trajectory diagram.

According to the existing drilling hydraulics theory (Singh *et al.*, 2021; Miao *et al.*, 2023), the fluid flow velocity in the eccentric annulus will not have a standard circular velocity distribution, and the overall flow velocity at the wide gap of the eccentric annulus flow field is larger than that at the narrow gap. Velocity is lower at the drill pipe and well wall contact position, and higher at the center. The flow rate at the drill pipe and well wall is zero, which is consistent with the assumption of no slip at the wall boundary. Let us take the working condition of 20 % eccentricity, 8 L/s displacement and 60 rev/min as an example. Let us create a cross-section at 12 m from the model axis, in this case, a 73 mm drill pipe cross-section from the drilling tool set. The axial flow velocity cloud at this cross-section is shown in Fig. 5. Let us create a line from the narrow gap to the wide gap at each of the 0.1 m (bit), 5 m (MWD), 12 m (drill pipe), and 15 m

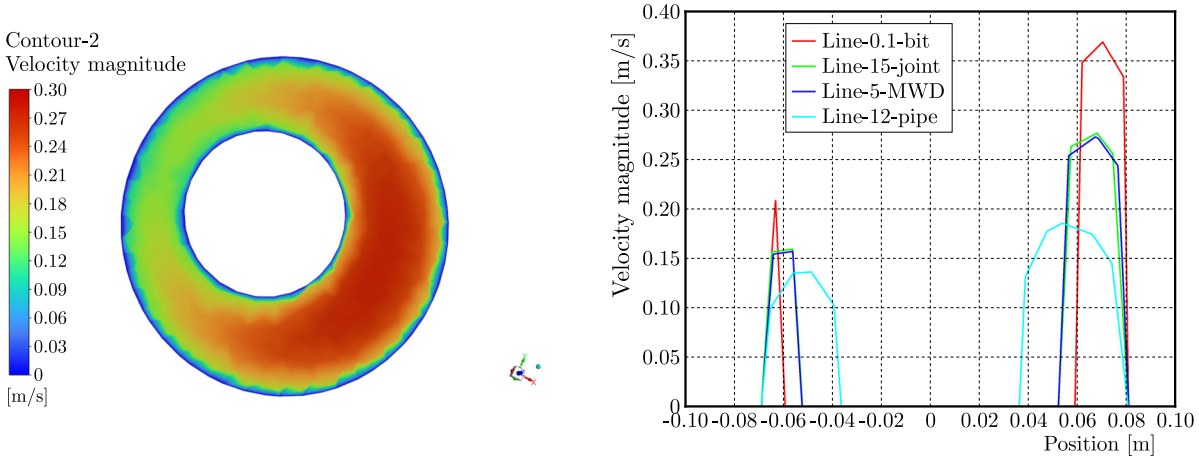

Fig. 5. Flow velocity cloud chart and distribution along the line at 12 m cross-section.

(joint) sections. The distribution of the flow velocity extending the line is obtained and plotted, and the results are shown in Fig. 5.

During the drilling fluid circulation process, the pressure distribution law of the flow field in the annulus can be basically expressed as the fluid dynamic pressure distribution law. The dynamic pressure distribution of the fluid in the annulus is similar to the velocity distribution, showing the phenomenon of low dynamic pressure at the position of fluid contacting the drill pipe and the well wall, and high dynamic pressure in the middle of the gap, and the dynamic pressure at the wide gap is larger than that at the narrow gap. The dynamic pressure cloud at 12 m section and the distribution of dynamic pressure along the line at 0.1 m (drill bit), 5 m (MWD), 12 m (drill pipe), and 15 m (joint) sections are shown in Fig. 6.

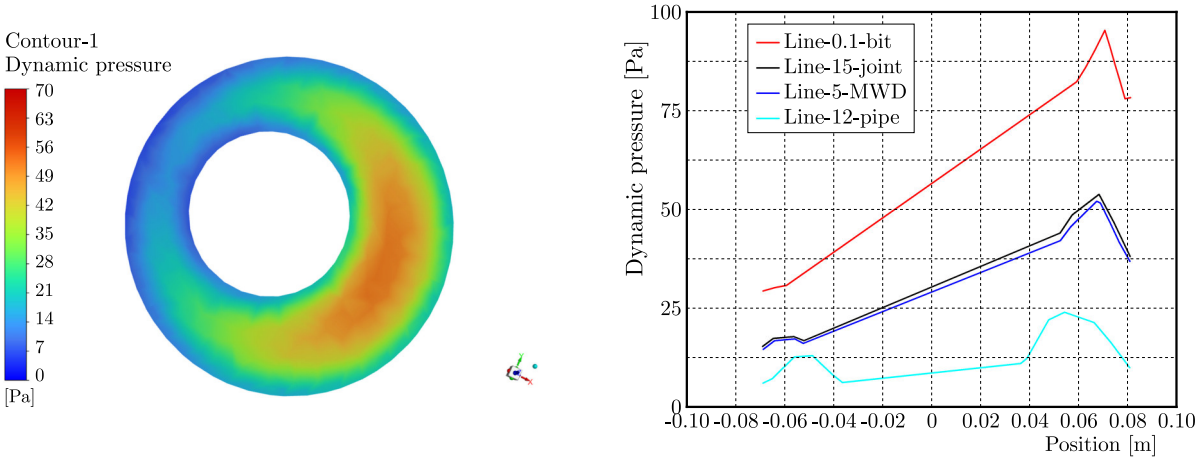

Fig. 6. Dynamic pressure cloud chart and distribution along the line at 12 m cross-section.

3.2. Results of numerical simulation of pressure drop gradient in the annulus

Let us analyze the variation of static pressure along the z-axis of the fluid domain model. Let us take the working condition of 20 % eccentricity, 8 L/s displacement and 60 rev/min as an example. The simulation results for the variation of drilling fluid static pressure along the z-axis of the fluid domain model are shown in Fig. 7. It can be seen that the gradient of pressure drop in the annulus is significantly different at different combinations of drilling tools. The pressure drop gradient is maximum at the drill bit with 560.6166 Pa/m and minimum at the drill pipe with 72.1636 Pa/m.

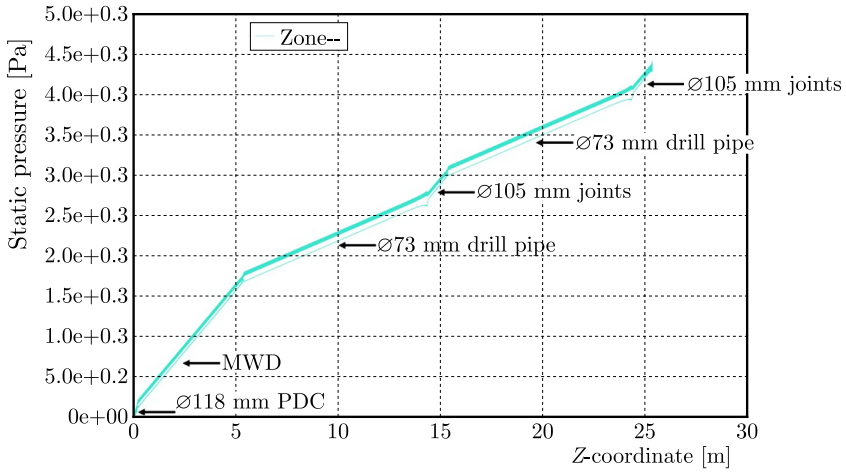

Fig. 7. Static pressure changes along the z-axis.

Let us simulate different speed conditions under different displacement for different eccentricity models. Let us record the pressure difference variation along the z-axis of the fluid domain model and calculate the pressure drop gradient. The calculation results of the annular pressure drop gradient under different models and operating conditions are shown in Tables 1–5.

Table 1. Numerical simulation results of pressure drop gradient in eccentric rotating annulus under a displacement of 5 L/s.

Eccentricity [%]	Gradient of annular pressure drop at different speeds [Pa/m]				
	0 rev/min	30 rev/min	60 rev/min	90 rev/min	120 rev/min
0	120.4681	120.5886	121.7639	115.3544	122.6413
10	117.4402	117.9311	118.7662	113.5687	118.0307
20	113.8847	114.4879	115.4942	111.7446	114.1889
30	110.5325	111.5087	112.6088	110.8481	111.8028
40	105.0750	105.7200	106.8325	104.0990	106.5690

Table 2. Numerical simulation results of pressure drop gradient in eccentric rotating annulus under a displacement of 8 L/s.

Eccentricity [%]	Gradient of annular pressure drop at different speeds [Pa/m]				
	0 rev/min	30 rev/min	60 rev/min	90 rev/min	120 rev/min
0	202.8120	203.2494	204.5276	206.7835	210.6868
10	196.9904	197.6563	199.2859	202.2209	206.8817
20	192.0572	192.8892	195.1042	198.7576	202.8354
30	183.4290	184.2530	186.3566	190.2460	192.3735
40	202.8120	203.2494	204.5276	206.7835	210.6868

Table 3. Numerical simulation results of pressure drop gradient in eccentric rotating annulus under a displacement of 11 L/s.

Eccentricity [%]	Gradient of annular pressure drop at different speeds [Pa/m]				
	0 rev/min	30 rev/min	60 rev/min	90 rev/min	120 rev/min
0	289.9956	290.5888	292.2413	304.2785	308.9576
10	292.2724	292.8492	294.4537	297.2361	302.2488
20	284.2837	285.2724	287.4275	290.9739	296.9511
30	278.0806	279.5182	282.2917	286.9200	293.9188
40	266.3290	267.9681	270.9574	275.6357	283.1700

Table 4. Numerical simulation results of pressure drop gradient in eccentric rotating annulus under a displacement of 14 L/s.

Eccentricity [%]	Gradient of annular pressure drop at different speeds [Pa/m]				
	0 rev/min	30 rev/min	60 rev/min	90 rev/min	120 rev/min
0	477.7297	478.2641	479.8586	482.7228	487.5286
10	465.3709	466.2607	467.9278	471.7419	476.8852
20	453.3478	455.1508	457.9864	462.7643	468.4593
30	445.0284	447.9129	452.0198	458.6739	466.1605
40	427.8005	431.2983	436.0922	443.1240	452.7278

Table 5. Numerical simulation results of pressure drop gradient in eccentric rotating annulus under a displacement of 17 L/s.

Eccentricity [%]	Gradient of annular pressure drop at different speeds [Pa/m]				
	0 rev/min	30 rev/min	60 rev/min	90 rev/min	120 rev/min
0	596.7309	597.3693	599.2548	602.1814	607.0979
10	581.4885	582.485	584.6119	588.5206	593.8054
20	507.1162	569.0454	572.3498	577.2255	584.1269
30	557.3839	560.9270	566.0114	573.1992	581.4801
40	536.7739	541.2815	547.4001	555.9908	563.7933

3.3. Three-factor analysis of eccentricity, rotation, and displacement

Let us draw a four-dimensional scatter plot by combining different eccentricity models, displacement, and rotation settings with data from Tables 1–5. In Fig. 8, the x-axis represents rotational speed, y-axis represents displacement, and z-axis represents eccentricity. The scattered points' color represents the magnitude of the annular pressure drop gradient. From the scatter plot, it can be clearly seen that the pressure drop gradient in the annulus tends to increase with the increase in speed and displacement. Combining the scatter plot matrix in Fig. 9, the histogram at the diagonal position allows us to see the distribution of each variable, while the scatter plots above and below the diagonal show the relationship between variables pairwise. There is a significant monotonic relationship between displacement and the annular pressure drop gradient. The pressure drop gradient in the annulus is more significantly affected by the rotational speed at high speeds (>90 rev/min) than at low speeds. As the eccentricity increases, the pressure drop gradient in the annulus slightly decreases.

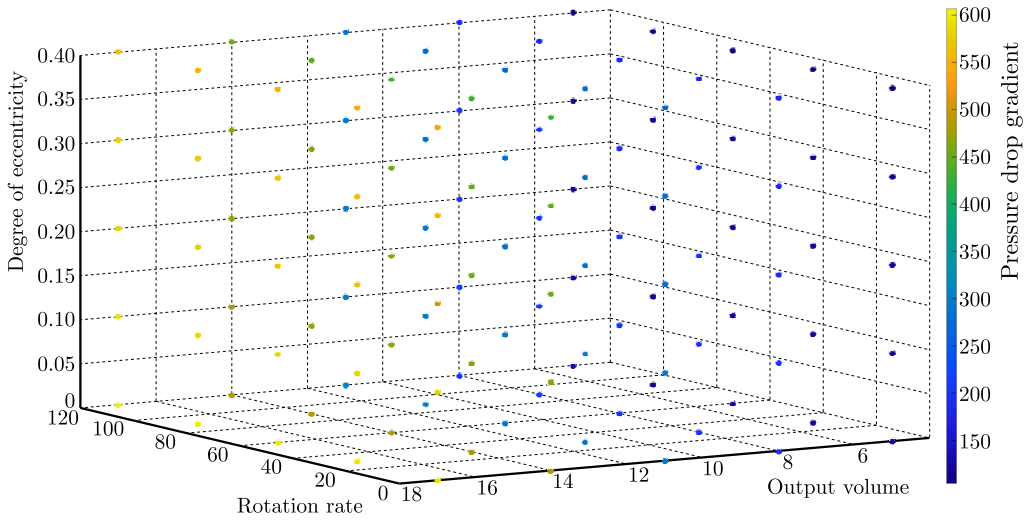

Fig. 8. Four-dimensional scatter plot analysis of three factors.

We used the coefficient comparison method in a multiple regression model to analyze the magnitude of the gradient effect of eccentricity, rotation, and displacement on the annular pressure drop (Bao & Weng, 2000; Mielke & Berry, 2002). Usually, while constructing a multifactor regression model, the equations are presented with unstandardized regression coefficients. It is the original regression coefficient corresponding to different independent variables in the equation, reflecting the magnitude of the effect of each unit change in the independent variable on the dependent variable when other factors remain constant. The variables in this article have asynchronous changes during simulation experiments. Let us standardize variables when

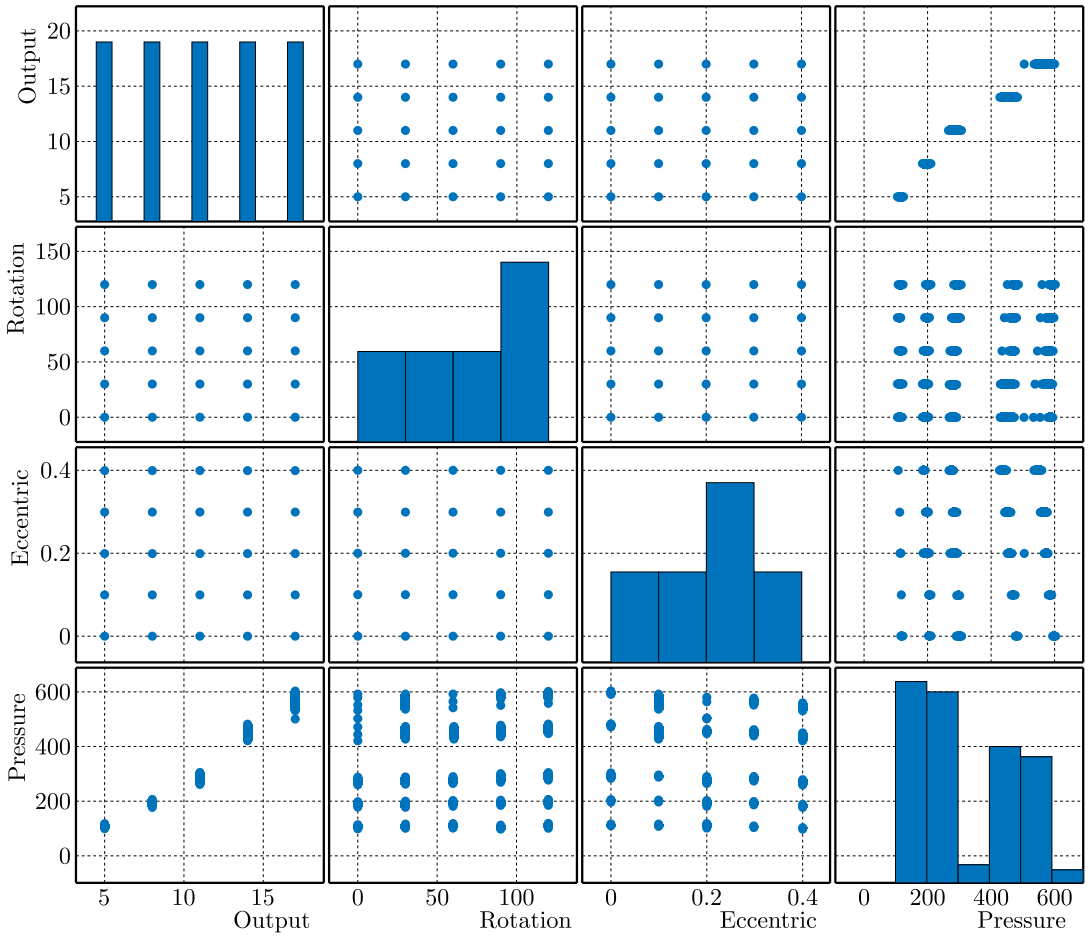

Fig. 9. Scatter plot matrix.

incorporating them into the regression model. At this point, it reflects the impact of every standard deviation change in the independent variable on the dependent variable. The standardized coefficient regression results are shown in Table 6. Through coefficient comparison, it can be seen that the effect of displacement on the annular pressure drop gradient is the most significant, while the effect of speed and eccentricity on the annular pressure drop gradient is relatively small, and eccentricity is negatively correlated with the annular pressure drop gradient.

Table 6. Normalized coefficient regression.

Coefficient					
Model	Non standardized coefficient		Standardized coefficient	t	Significance
	B	Standard error	Beta		
Constant	−98.280	7.397	–	−13.287	1.000
Displacement	39.345	0.509	0.988	77.267	0.000
Speed	0.114	0.051	0.029	2.237	0.027
Eccentricity	−72.644	15.276	−0.061	−4.755	0.000
Dependent variable: pressure drop gradient					

Let us create a scatter plot of the annular pressure drop gradient with smooth lines and data labels for different speeds and eccentricities at various displacements. As shown in Fig. 10, when the displacement is less than 11 L/s and the speed is less than 60 rev/min, the pressure drop gradient in the annulus increases with the increase in speed. When the speed is greater than

60 rev/min, the pressure drop gradient in the annulus slightly decreases with increasing speed. When the speed increases to 90 rev/min or above, the pressure drop gradient in the annulus continues to rise with the increase in speed. When the displacement is ≥11 L/s, the annular pressure drop gradient increases monotonically with the increase in rotational speed, and the higher the rotational speed, the greater the impact on the annular pressure drop gradient. When the eccentricity is less than 0.3 and the displacement is greater than 14 L/s, the pressure drop gradient in the annulus at low speed initially increases and then decreases with the increase in eccentricity. The overall trend is that the annular pressure drop gradient decreases with increasing eccentricity, and the larger the eccentricity, the greater the impact on the annular pressure drop gradient. Among them, when the displacement is 5 L/s and the speed is 90 rev/min, all eccentricity model annular pressure drop gradients show a decreasing trend. The eccentricity 0.4 model achieves the lowest annular pressure drop gradient at this displacement at this speed. At a displacement of 11 L/s, the annular pressure drop gradient of the model with 0 eccentricity is smaller than that of the model with 0.1 eccentricity before reaching 60 rev/min. At speeds of 90 rev/min and above, the annular pressure drop gradient of the model with 0 eccentricity increases faster, surpassing the model with 0.1 eccentricity. At a displacement of 17 L/s, the model with an eccentricity of 0.2 obtains the minimum annular pressure drop gradient at a speed of 0 rev/min.

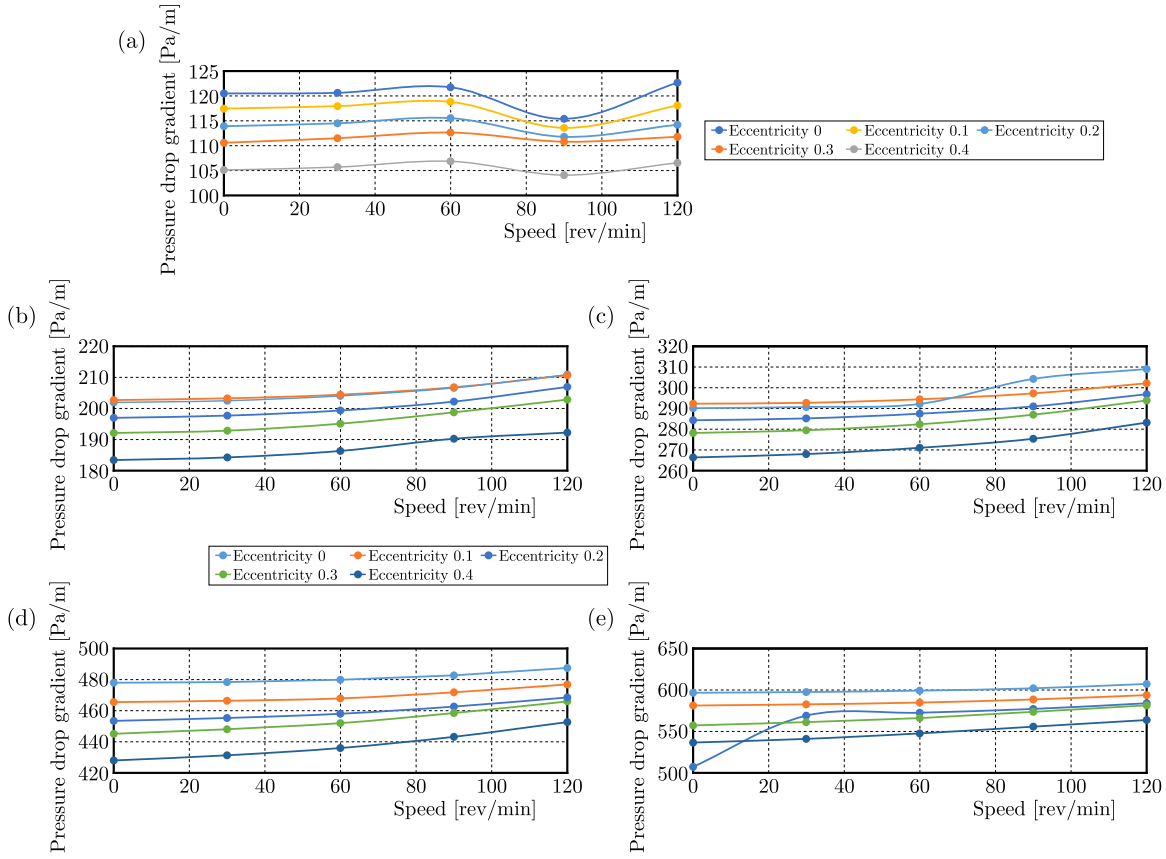

Fig. 10. Pressure drop gradient in the annulus under different speeds and eccentricities with displacements of: (a) 5 L/s; (b) 8 L/s; (c) 11 L/s; (d) 14 L/s; (e) 17 L/s.

3.4. Differences in pressure drop gradients in the annulus for different combinations of drilling tools

Let us create a pipe string model through Solidworks to explore the annular pressure loss of different drilling tool combinations. Using eccentricity 0 and eccentricity 0.1 models as simulation experimental objects, let us record the pressure difference between the upper and lower sections of

each drilling tool, and obtain the annular pressure drop gradient. Let us analyze the differences in the annular pressure drop gradient at different drilling tool combinations under different displacements, eccentricities, and speeds. The simulation experiments are shown in Tables 7–11.

Table 7. Calculation results of annular pressure drop gradient at different drilling tools under 5 L/s displacement.

Different eccentricity models	Different simulated drilling tools	The pressure drop gradient in the annulus at different rotational speeds [Pa/m]				
		0 rev/min	30 rev/min	60 rev/min	90 rev/min	120 rev/min
Eccentricity 0	Simulated drill bit	577.2046	566.7762	545.4048	527.0940	509.5932
	Simulated MWD	234.1352	232.7951	231.9452	230.0234	228.5605
	Simulated joint	239.2424	241.3771	248.4087	265.7412	288.5091
	Simulated drill pipe	72.6384	72.8911	73.4337	65.8288	73.8904
Eccentricity 0.1	Simulated drill bit	568.2580	560.6166	541.1510	514.6571	510.2580
	Simulated MWD	246.8312	224.2789	222.2801	219.8526	218.8613
	Simulated joint	231.0896	234.1507	239.8133	257.5202	276.3035
	Simulated drill pipe	71.8530	72.1636	72.7504	66.2534	70.37980

Table 8. Calculation results of annular pressure drop gradient at different drilling tools under 8 L/s displacement.

Different eccentricity models	Different simulated drilling tools	The pressure drop gradient in the annulus at different rotational speeds [Pa/m]				
		0 rev/min	30 rev/min	60 rev/min	90 rev/min	120 rev/min
Eccentricity 0	Simulated drill bit	980.4990	971.7729	946.0966	908.8517	866.2889
	Simulated MWD	393.8126	393.3917	392.1140	390.5879	388.8704
	Simulated joint	401.6579	404.1619	411.9381	426.0342	449.4999
	Simulated drill pipe	121.6249	121.7959	122.8887	124.7786	129.9980
Eccentricity 0.1	Simulated drill bit	1000.1140	992.7490	970.9457	940.6719	913.6777
	Simulated MWD	377.1419	376.7523	375.4036	372.9664	373.2648
	Simulated joint	392.2802	394.2948	401.0442	413.8018	432.2522
	Simulated drill pipe	127.7627	128.0762	129.0871	131.9266	133.1912

Table 9. Calculation results of annular pressure drop gradient at different drilling tools under 11 L/s displacement.

Different eccentricity models	Different simulated drilling tools	The pressure drop gradient in the annulus at different rotational speeds [Pa/m]				
		0 rev/min	30 rev/min	60 rev/min	90 rev/min	120 rev/min
Eccentricity 0	Simulated drill bit	1486.7360	1480.1770	1459.6640	1428.7510	1394.7340
	Simulated MWD	535.3307	535.1713	534.5935	533.6185	533.1697
	Simulated joint	560.3379	560.8369	569.6386	582.5776	604.7408
	Simulated drill pipe	186.9023	187.2892	188.5602	190.5277	194.7520
Eccentricity 0.1	Simulated drill bit	1455.2450	1447.1880	1422.6190	1427.3020	1383.2950
	Simulated MWD	559.1121	558.8473	557.9023	556.9548	556.3359
	Simulated joint	573.2401	576.1122	584.0520	602.3688	624.3576
	Simulated drill pipe	176.7335	176.9028	178.1618	191.8827	195.6730

Table 10. Calculation results of annular pressure drop gradient at different drilling tools under 14 L/s displacement.

Different eccentricity models	Different simulated drilling tools	The pressure drop gradient in the annulus at different rotational speeds [Pa/m]				
		0 rev/min	30 rev/min	60 rev/min	90 rev/min	120 rev/min
Eccentricity 0	Simulated drill bit	2535.1230	2528.229	2506.491	2472.6240	2425.1050
	Simulated MWD	881.3331	881.1474	880.4879	880.0918	879.4367
	Simulated joint	920.9313	923.4875	930.8826	942.2473	962.8600
	Simulated drill pipe	303.7599	304.1637	305.1402	306.6544	310.4984
Eccentricity 0.1	Simulated drill bit	2524.9240	2520.2230	2501.6560	2473.846	2437.0160
	Simulated MWD	842.1706	842.5658	842.0543	843.0256	843.4042
	Simulated joint	889.7953	892.5088	899.2022	912.9028	933.2046
	Simulated drill pipe	300.3946	300.9917	302.2901	305.2190	309.4928

Table 11. Calculation results of annular pressure drop gradient at different drilling tools under 17 L/s displacement.

Different eccentricity models	Different simulated drilling tools	The pressure drop gradient in the annulus at different rotational speeds [Pa/m]				
		0 rev/min	30 rev/min	60 rev/min	90 rev/min	120 rev/min
Eccentricity 0	Simulated drill bit	3297.4322	3291.0201	3271.5220	3238.6260	3194.9181
	Simulated MWD	1096.8991	1096.7751	1096.8383	1096.3944	1096.0871
	Simulated joint	1151.0115	1153.6450	1161.0901	1172.8780	1193.5186
	Simulated drill pipe	380.0268	380.6023	381.8937	383.9324	388.3010
Eccentricity 0.1	Simulated drill bit	3286.5050	3282.4980	3266.0181	3240.2089	3206.3397
	Simulated MWD	1048.6480	1049.1877	1049.0842	1050.1278	1052.3933
	Simulated joint	1112.9489	1115.7730	1123.3630	1137.1562	1156.3222
	Simulated drill pipe	375.9475	376.6142	378.4151	381.8080	384.9924

According to the simulated calculation data, it can be seen that the annular pressure drop gradient at the drill bit is the largest and the annular pressure drop gradient at the drill rod is the smallest. This is because the size of the drill bit is larger than that of the drill rod, resulting in a smaller annular clearance, faster flow velocity, and higher dynamic pressure on the cross-section. However, in actual drilling, due to the overall length of the drill pipe being much longer than the drill bit, the main annular pressure loss is still distributed at the drill pipe. The size set by MWD in the simulation is larger than that of the drilling tool joint. The pressure drop gradient in the annulus at the MWD is greater than that at the joint of the drilling tool only in the 0.1 eccentricity model, when simulating working conditions at 5 L/s and 0 rev/min speed. In the overall trend, the pressure drop gradient in the annulus at the MWD is smaller than that in the drilling tool joint. Moreover, as the displacement and speed increase, the difference becomes greater. This is mainly because the MWD length is greater than that of the drilling tool joint, and the diameter change of the drilling tool near the MWD is smaller, while the diameter change at the drilling tool joint is more obvious, and the fluid is more affected. This indicates that in actual drilling, the impact of drilling tool joints or sudden diameter changes on annular pressure loss is difficult to ignore. Let us explore the influence of drilling tool joints on annular pressure loss based on the 0.1 eccentricity model. Let us simulate the difference in the annular pressure drop gradient with and without joints at different displacements at 0 rev/min. The simulation results are shown in Fig. 11. From the figure, it can be seen that the difference in the annular pressure drop gradient between the joint and the non-joint is over 200 %,

and even more than three times at a displacement of 5 L/s. Furthermore, as the displacement increases, the impact of the joint on the annular pressure drop gradient becomes greater.

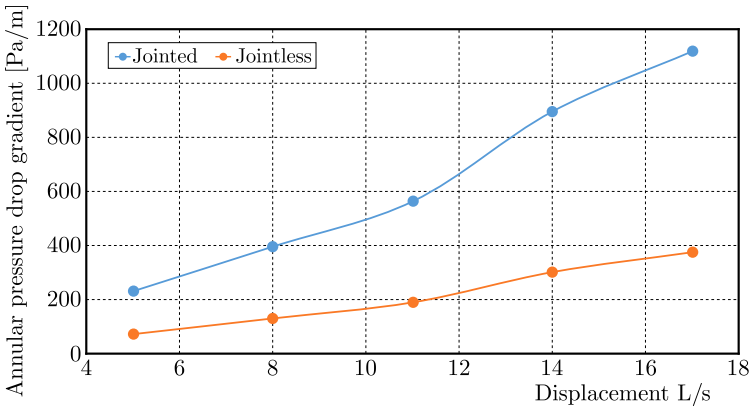

Fig. 11. Pressure drop gradient in annular space with and without joints at different displacements.

4. Fitting correction factors and validation

Let us divide the annular pressure drop gradient data in Tables 1–5 by the annular pressure drop gradient at a concentric non-rotating 8 L/s displacement to obtain the dimensionless annular pressure drop gradient. Let us perform multivariate fitting to obtain the multivariate fitting formula:

$$\Delta P_N = -0.487 + 0.195V + 0.001R - 3.6E, \tag{4.1}$$

where ΔP_N is the dimensionless annular pressure drop gradient factor, V is the displacement, R is the speed, and E is the eccentricity.

Based on the historical data of the SY-3 well, a horizontal section of the SY-3 well was selected as the calculation point for lateral drilling. The drilling tool combination consisted of ∅118 mm PDC + ∅95 mm Screw drilling tools (1.5°) + ∅103 mm non-magnetic drill pipe + directional joint + 103 mm non-magnetic drill pipe + ∅73 mm weighted drill pipe + hydraulic oscillator + ∅73 mm weighted drill pipe + ∅73 mm drill pipe + ∅73 mm weighted drill pipe + ∅73 mm drill pipe. Drilling fluid density is 1400 kg/m^3, viscosity is 0.058 Pa · s, displacement is 7 L/s, speed is 40 rev/min. At this time, the pump pressure is 23 Mpa. Based on the rheological properties of drilling fluid and the size of the flow channel, the pressure drop gradient ΔP of the H-B fluid annulus under concentric non-rotating 8 L/s displacement is obtained through an analytical formula. Based on the characteristics of the wellbore section such as the inclination angle, slope angle, and drill string size, the eccentricity E is determined using the drill string buckling theory (Vaughn, 1965; Juvkam-Wold & Wu, 1992; Lubinski & Althouse, 1962; Dawson, 1984; Tian *et al.*, 2024). Let us calculate the dimensionless annular pressure drop gradient factor based on Eq. (4.1) for displacement V and speed R, and correct ΔP. The system's cyclic pressure loss was calculated based on the well history data. Besides, the error was less than 10 % when compared with the annular pressure loss calculated by the fitting model. This validates the accuracy of the model.

5. Conclusions

– After analyzing the hydraulic behavior of eccentric rotating annular fluid in a small wellbore through Fluent simulation, the overall flow velocity at the wide gap of the eccentric annular flow field was found to be greater than that at the narrow gap. The speed is lower at the contact position between the drill pipe and the wellbore wall, but higher at the center. The dynamic pressure distribution and velocity distribution characteristics of the fluid

in the annulus are similar, showing a phenomenon of low dynamic pressure at the contact position between the fluid and the drill pipe and wellbore while high dynamic pressure is observed in the middle of the gap.

- The gradient of annular pressure drop was obtained through simulation calculation results. A comprehensive analysis was conducted on the three factors affecting annular pressure loss, namely eccentricity, rotation, and displacement. At low speeds, the annular pressure drop gradient first increases and then decreases with increasing eccentricity, but overall there is a trend of the annular pressure drop gradient decreasing with increasing eccentricity. The greater the eccentricity, the greater the impact of eccentricity on the pressure drop gradient in the annulus.
- The annular pressure drop gradient was simulated and calculated with different drilling tool combinations. The differences were analyzed in the annular pressure drop gradient at different drilling tool combinations under different displacements, eccentricities, and speeds. The impact of drilling tool joints on annular pressure loss was explored. As the displacement increases, the impact of the joint on the annular pressure drop gradient becomes greater.
- Multi factor fitting of the dimensionless annular pressure drop gradient factor was analyzed through numerical simulation results. The pump pressure of a certain lateral drilling horizontal section of SY-3 well was analyzed and combined with well history data. The prediction error is less than 10 %, which verifies the accuracy of the model.

References

1. Bao, F.D., & Weng, X.H. (2000). The software solving of multiple regression and correlation analysis and case explanation (in Chinese). *Journal of Applied Statistics and Management*, *20*(5), 56–61.
2. Cartalos, U., King, I., Dupuis, D., & Sagot, A. (1996). Field validated hydraulic model predictions give guidelines for optimal annular flow in slimhole drilling. In *IADC/SPE Drilling Conference* (Article SPE-35131-MS). Society of Petroleum Engineers. https://doi.org/10.2118/35131-MS
3. Dawson, R. (1984). Drill pipe buckling in inclined holes. *Journal of Petroleum Technology*, *36*(10), 1734–1738. https://doi.org/10.2118/11167-PA
4. Delwiche, R.A., Lejeune, M.W.D., Mawet, P.F.B.N., & Vignette, R. (1992). Slimhole drilling hydraulics. In *SPE Annual Technical Conference and Exhibition* (Article SPE-24596-MS). Society of Petroleum Engineers. https://doi.org/10.2118/24596-MS
5. Enfis, M., Ahmed, R., & Saasen, A. (2011). The hydraulic effect of tool-joint on annular pressure loss. In *SPE Production and Operations Symposium* (Article SPE-142282-MS). Society of Petroleum Engineers. https://doi.org/10.2118/142282-MS
6. Haciislamoglu, M., & Cartalos, U. (1994). Practical pressure loss predictions in realistic annular geometries. In *SPE Annual Technical Conference and Exhibition* (Article SPE-28304-MS). Society of Petroleum Engineers. https://doi.org/10.2118/28304-MS
7. Hansen, S.A., Rommetveit, R., Sterri, N., Aas, B., & Merlo, A. (1999). A new hydraulics model for slim hole drilling applications. In *SPE/IADC Middle East Drilling Technology Conference* (Article SPE-57579-MS). Society of Petroleum Engineers. https://doi.org/10.2118/57579-MS
8. Hemphill, T., & Ravi, K. (2005). Calculation of drillpipe rotation effects on fluids in axial flow: An engineering approach. In *SPE Annual Technical Conference and Exhibition* (Article SPE-97158-MS). Society of Petroleum Engineers. https://doi.org/10.2118/97158-MS
9. Juvkam-Wold, H.C., & Wu, J. (1992). Casing deflection and centralizer spacing calculations. *Society of Petroleum Engineers Drilling Engineering*, *7*(4), 268–274. https://doi.org/10.2118/21282-PA
10. Kelessidis, V.C., Dalamarinis, P., & Maglione, R. (2011). Experimental study and predictions of pressure losses of fluids modeled as Herschel–Bulkley in concentric and eccentric annuli in laminar, transitional and turbulent flows. *Journal of Petroleum Science and Engineering*, *77*(3–4), 305–312. https://doi.org/10.1016/j.petrol.2011.04.004

11. Khatibi, M., Wiktorski, E., Sui, D., & Time, R.W. (2018). Experimental study of frictional pressure loss for eccentric drillpipe in horizontal wells. In *IADC/SPE Asia Pacific Drilling Technology Conference and Exhibition* (Article SPE-191046-MS). Society of Petroleum Engineers. https://doi.org/10.2118/191046-MS

12. Letelier, M.F., Siginer, D.A., & Hinojosa, C.B. (2017). On the physics of viscoplastic fluid flow in non-circular tubes. *International Journal of Non-Linear Mechanics*, *88*, 1–10. https://doi.org/10.1016/j.ijnonlinmec.2016.09.012

13. Lubinski, A., & Althouse, W.S. (1962). Helical buckling of tubing sealed in packers. *Journal of Petroleum Technology*, *14*(6), 655–670. https://doi.org/10.2118/178-PA

14. McCann, R.C., Quigley, M.S., Zamora, M., & Slater, K.S. (1995). Effects of high-speed pipe rotation on pressures in narrow annuli. *Society of Petroleum Engineers Drilling & Completion*, *10*(2), 96–103. https://doi.org/10.2118/26343-PA

15. Miao, H., Dokhani, V., Ma, Y., & Zhang, D. (2023). Numerical modeling of laminar and turbulent annular flows of power-law fluids in partially blocked geometries. *Results in Engineering*, *17*, Article 100930. https://doi.org/10.1016/j.rineng.2023.100930

16. Mielke, P.W., & Berry, K.J. (2002). Multivariate multiple regression analyses: A permutation method for linear models. *Psychological Reports*, *91*(1), 3–9. https://doi.org/10.2466/pr0.2002.91.1.3

17. Reed, T.D., & Pilehvari, A.A. (1993). A new model for laminar, transitional, and turbulent flow of drilling muds. In *SPE Production Operations Symposium* (Article SPE-25456-MS). Society of Petroleum Engineers. https://doi.org/10.2118/25456-MS

18. Resell, Å.Aa., Giljarhus, K.E.T., Mihai, R., & Skadsem, H.J. (2025). Fluid forces in eccentric annular geometries with rotating and orbiting inner cylinder. *Physics of Fluids*, *37*(4), Article 047158. https://doi.org/10.1063/5.0262548

19. Shi, K., & Zhang, S. (2025). Turbulent transport in annular Poiseuille flow with axial rotation. *Physics of Fluids*, *37*(3), Article 035116. https://doi.org/10.1063/5.0257318

20. Singh, R., Ahmed, R., Karami, H., Nasser, M., & Hussein, I. (2021). CFD analysis of turbulent flow of power-law fluid in a partially blocked eccentric annulus. *Energies*, *14*(3), Article 731. https://doi.org/10.3390/en14030731

21. Song, Z.C., Wang, G.C., Guan, Z.C., & Zou, D.Y. (2004). A method for computing the circulating pressure loss in slim hole annulus (in Chinese). *Petroleum Drilling Techniques*, *32*(6), 11–12.

22. Sotoudeh, S., Frigaard, I.A. (2024). Computational study of Newtonian laminar annular horizontal displacement flows with rotating inner cylinder. *Physics of Fluids*, *36*(8), Article 083113. https://doi.org/10.1063/5.0222314

23. Tian, J., Song, H., Yang, Y., Mao, L., & Song, J. (2024). Dynamic buckling characteristics of drill string in horizontal wells. *International Journal of Structural Stability and Dynamics*, *24*(11), Article 2450121. https://doi.org/10.1142/S0219455424501219

24. Tian, Y., Jiang, D.L., Ma, C.H., Xu, Y.L., Yu, X.D., & Song, X.C. (2022). Numerical simulation of the effects of eccentric rotation of the drill string on annular frictional pressure drop (in Chinese). *Petroleum Drilling Techniques*, *50*(5), 42–49. https://doi.org/10.11911/syztjs.2022104

25. Vaughn, R.D. (1965). Axial laminar flow of non-Newtonian fluids in narrow eccentric annuli. *Society of Petroleum Engineers Journal*, *5*(4), 277–280. https://doi.org/10.2118/1138-PA

26. Vieira Neto, J.L., Martins, A.L., Ataíde, C.H., & Barrozo, M.A.S. (2014). The effect of the inner cylinder rotation on the fluid dynamics of non-Newtonian fluids in concentric and eccentric annuli. *Brazilian Journal of Chemical Engineering*, *31*(4), 829–838. https://doi.org/10.1590/0104-6632.20140314s00002871

Manuscript received September 17, 2024; accepted for publication July 31, 2025; published online October 20, 2025.

Contents

Walczak T., Guminiak M., Kamiński M., *Probabilistic estimation of the dynamic gait parameters* 715

Ciurej H., *Sensitivity analysis of multiple eigenvalues and associated eigenvectors of quadratic eigenproblem* 721

Knitter-Piątkowska A., Kawa O., Guminiak M., *Damage localization in the main structural elements of steel halls applying dynamic structural response signal and discrete wavelet transform* 737

Magnucki K., *Elastic buckling of an individual I-beam with consideration of the shear effect* 743

Kurpanik K., Hartwich J., Kciuk S., Duda S., *Impact acceleration acquisition with a high frequency data logging system based on SPI communication protocol* 755

Brzozowski M., *The model-based algorithm for autonomous vehicle path following* 769

Smyk E., Stopel M., Rachwalski A., *Impact of honeycomb straightener parameters on operation in a straight duct* 781

Prażmowski M., Małecka J., Łagoda T., *Effect of fatigue on the microhardness of scrap cross-sections after cyclic bending with torsion of RG7 bronze alloy* 795

Jereczek B., Maciejewski I., Błażejewski A., Pecolt S., Krzyżyński T., *Research on the energy recovery system in the active horizontal seat suspension of a working machine* 809

Mazurkiewicz A., Pezowicz C., Nikodem A., Przybyłek M., Mazurkiewicz D., *Tribological effects of hyaluronic acid concentration on articular cartilage: Pin-on-plate friction tests in porcine and osteoarthritic human tissue* 815

Zhou M., Zhong J., Zhang Q., Gan Y., Zhang C., Liu H., *Investigation of flow control on a vertical axis wind turbine using a bionic flap* 823

He Q., Yuan Q., Lang L., *A fast and robust numerical solution for the positioning design of horizontal drains in a saturated-unsaturated soil ground system* 839

Jia M., Lan X., Zhuang T., Sun B., *The impact of racket string tension on tennis racket performance* 855

Liu C., Li L.J., Hu S.W., *Influence of conical structure on sealing specific pressure under static loading* 865

Yu P., Sun W., Suo Y., Wu Y., *Analysis of mechanochemical diffusion coupling processes based on transient continuum chemo-mechanical coupling theory* 877

Qin Z., *Experimental investigation of the long-term mechanical behavior of mudstone under varying water contents* 889

Jędrejko P., *Turbulent coherent structures in thermal vortex rings* 903

Wang Z., Shen Y., Hou Y., Kan Y., Huang W., Zhu Y., *Simulated calculation and application of annular pressure loss for deep slim-hole sidetracking horizontal well* 915