**Proceedings of the Design Society, Volume 5: ICED25**
https://doi.org/10.1017/pds.2025.10281

# Quantitative metrics for validation and decision-making in digital twins: a comparative study on a railway braking system

**Dmitrii Ershenko ✉, Glafira Derbysheva, Andreas Panayi and Clement Fortin**

*Skolkovo Institute of Science and Technology (Skoltech), Russia*

✉ dmitrii.ershenko@skoltech.ru

**ABSTRACT:** The overall quality of final Digital Twin (DT) solutions and their ability to produce useful insights are key considerations for researchers and for the industry to readily adopt them. However, validation of DTs is often neglected in existing research dedicated to their development. Further, there is a lack of methodologies for building bi-directional information exchanges between virtual and real spaces, potentially hindering effective decision-making. This work presents a comparative analysis of several quantitative metrics by implementing them on the Digital Twin of a railway braking system as a use case. Their suitability as performance measures for validation and as thresholds to support decision-making is assessed. Their integration into a novel DT structure is shown to contribute to a well-rounded validation procedure and a practical decision-making framework.

**KEYWORDS:** digital twin, simulation, decision making, functional modelling, validation

## 1. Introduction

### 1.1. Digital twin

Historically, numerous attempts to reproduce the behavior of physical engineering systems in the virtual domain have been made, setting the stage for Digital Twins, most notably as part of the Apollo mission (Boschert & Rosen, 2016). However, the concept of the Digital Twin was not introduced until 2003. As part of the Executive Course on Product Lifecycle Management (PLM) at the University of Michigan, Dr. Michael Grieves proposed that, by definition, a Digital Twin consists of three parts: physical entity, virtual entity and bi-directional information flow between them, as illustrated in Figure 1 (Grieves, 2015).

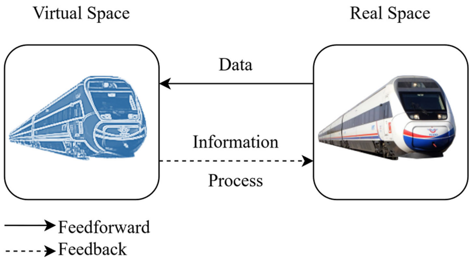

**Figure 1. DT as a three parts concept**

As the number and variety of DT applications increased, efforts have been made to expand the definition to adequately account for these varying applications and to develop standardized frameworks for the development of DTs. Consequently, the concept of the Structured Traceable Efficient and Manageable (STEM) Digital Twin was introduced in a predecessor study to the current paper (Ershenko et al., 2024).

This model is a combination of several ideas and aims to provide a standardized approach for the development of DTs.

As shown in Figure 2, the STEM model includes the three classic elements of a DT, i.e. virtual space, real space, and the feedback/feedforward data connection between them, as well as additional constructs defining the information flow and detailed internal structure of the virtual space. The design of the virtual space follows the SAPPhIRE methodology, which is a structured approach to describing system functionality through causally linked components (Chakrabarti et al., 2004; Srinivasan & Chakrabarti, 2009). This approach allows users to design new systems more effectively or to gain insight into existing ones by specifying the interactions and cause-effect relationships inside the system.

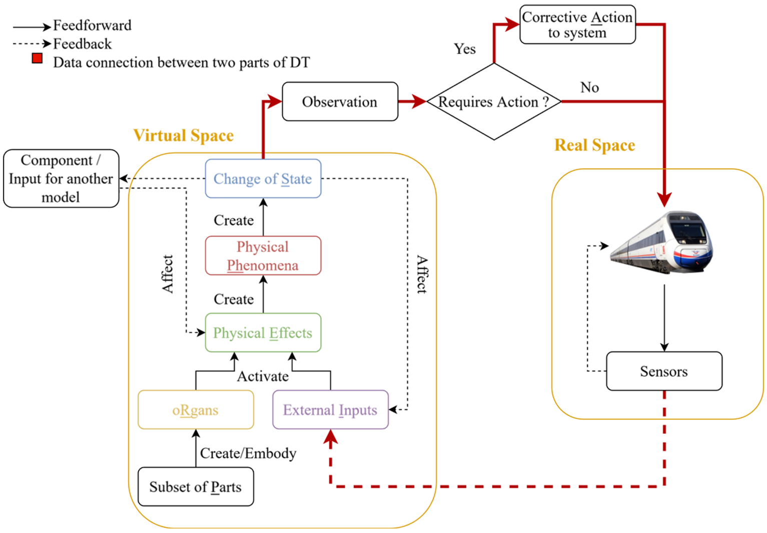

Figure 2. STEM digital twin

The original SAPPhIRE model defines *Action* as an interpretation of the *Change of State*, which is caused by the *Physical Phenomena* (Bhattacharya & Chakrabarti, 2023). From Figure 2, the feedforward information flow from the virtual space to the real space in the STEM model is facilitated by a dedicated process. The purpose of this process is to determine, based on an *Observation*, whether corrective *Action* is needed in the real space. The concept of an *Observation* was introduced as an extension of the original SAPPhIRE concept, to bridge the gap between *Change of State* and *Action*. Introduction of the *Observation* creates an unambiguous space, where the logic for interpreting the *Change of State* can be implemented and handled separately from the *Change of State* and the *Action* (McSorley et al., 2014). In practice, the logic of the *Observation* will differ depending on the system and the phenomena being monitored, the stage of the product lifecycle where the Digital Twin is implemented, e.g. design or maintenance stage, and the type of observer making the assessment, e.g. human or automated.

The STEM approach proposes that the DT triggers corrective *Action* when the deviation between real and simulated outputs (*Change of State*) exceeds a predefined threshold (*Observation*). It follows that a practical implementation of a DT built using the STEM methodology, as discussed in this paper, requires objective quantitative criteria for comparing real and simulated data that can serve as this threshold.

## 1.2. Verification and validation

The introduction of Digital Twins to any phase of the product lifecycle of a physical asset is associated with added value. A wide variety of use cases can be found ranging from the aerospace industry (Li et al., 2022) to agriculture (Pylianidis et al., 2021). For instance, DTs allow to save physical resources and time that would otherwise be spent on physical prototyping during the design phase. In the product support phase, DTs can be used for system failure analysis as part of a predictive maintenance strategy (Grieves & Vickers, 2017).

To gain these and other benefits associated with the use of DTs, establishing trust in them is essential. In large, this depends on the ability of the virtual part of the DT to reliably represent the physical product. The use of Verification and Validation (V&V) is a widely accepted strategy to ensure the reliability of models in simulation, systems and software engineering (Grieves, 2023). Despite this, the topic of V&V

in Digital Twin development remains under-represented in research (Bitencourt et al., 2023; Muñoz et al., 2022).

In the context of Digital Twins, validation of its virtual representation is defined as the process to determine the ability *“of the simulation model to reasonably represent the real world from the perspective of its intended purpose”*. Meanwhile, verification refers to ensuring that *“the model implementation is correct according to previously agreed specifications and assumptions”* (Bitencourt et al., 2023).

Referring to other works in the area of V&V in simulation engineering, it is suggested that model validation consists of two steps: face validity checks and quantitative validation (Hua et al., 2022). The idea behind face validity checks is to establish the model’s realism, hence the approaches are generally subjective. On the other hand, quantitative validation is an objective procedure and aims to test the similarity between the outputs of the physical system and its digital representation, i.e. simulation model, using predefined performance measures. Therefore, this is the step that has the most potential for standardization.

In fact, having analyzed several formal definitions of a *“Digital Twin”*, it is proposed that in order to qualify as a DT a model must be similar to its physical asset (Emmert-Streib, 2023). The authors highlight the need for a “similarity measure or distance measure” to perform this assessment. In other words, the paper presents the need for validation using predefined criteria necessary to classifying a solution as a Digital Twin.

### 1.3. Decision-making

As discussed in the previous section, the use of Digital Twins is associated with added value. This value is created when a DT makes a decision, based on simulation results, and the decision has an effect on the business process involving the physical asset (West et al., 2021). However, the authors also point out that existing research on DT decision-making is limited. Furthermore, more research is needed to develop standardized methods for information exchanges within Digital Twins, which enable the decision-making (Ma et al., 2024).

Nevertheless, some examples of using quantitative metrics to make decisions regarding the asset’s condition exist (Villalonga et al., 2020). In this research, the errors between the real and simulated outputs of a DT of a manufacturing system are calculated to determine, based on a pre-established threshold, whether the system is operating normally. Using this information, the decisions made by the DT contribute to improved operational management and more efficient task scheduling within the production system.

It follows that comparing the outputs of the physical and virtual entities may be used to gain insight into the asset’s state and inform corrective *Actions* that the DT provides to the physical system. Moreover, using predefined metrics to perform the comparison offers a standardized approach to support decision-making of a Digital Twin.

## 2. Research objective

The literature review presented in section 1.2 identified a research gap related to the lack of standardized approaches to DT validation, particularly in terms of universal similarity measures which can serve as quality criteria. Meanwhile, section 1.3 identified a need for the development of standardized methods for information exchange and decision-making within DTs.

The objective of this paper is to perform a comparative study of several quantitative metrics as quality measures in validation of the virtual part of a Digital Twin and as thresholds that determine a Digital Twin’s feedforward decision-making in relation to its physical entity. This will be done by implementing them on the STEM Digital Twin of a railway braking system as a use case. The construction of the STEM Digital Twin and implementation of metrics will be performed using Simcenter Amesim, a 1D simulation tool. The results are expected to be valid for implementations using alternative software and tools.

# 3. Research methodology

## 3.1. Use case

A Digital Twin of a railway pneumatic braking system will be considered as a use case. There are two main compressed air lines which support this system: the Feed Line (FL) and the Brake Line (BL). Two compressors supply air to the system, charging pressure in FL and BL. Afterwards, the brakes are in a released condition and ready to operate. Upon command, air enters the brake cylinder and the brakes are applied. Therefore, system health can be assessed by monitoring the pressure in BL and FL.

While the Brake Line is exclusively used for braking, the Feed Line also supplies compressed air to other consumers, such as sand systems, air horns and whistles. Sand systems dispense sand onto the track in front of the train wheel, increasing friction between them. Air horns and whistles are warning devices. An example train presented in this study has 5 carriages. The control panel that provides inputs to the brake handle and the other systems is located in car 1. Both compressors are located in car 3. Each car has a FL pressure sensor and a BL pressure sensor. In total there are 15 sensors distributed across 5 carriages that are relevant to this Digital Twin. They measure pressure in BL and FL and indicate the status (i.e. active/inactive) of compressors, sand systems, air horns and whistles. This combination of sensors is adequate for representing the system. It is possible to include additional parameters if they are available. The location of the sensors is shown in Figure 3.

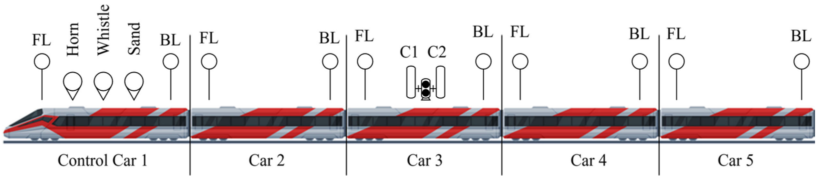

Figure 3. Train sensor distribution

## 3.2. Metrics

The choice of the quantitative validation procedure, which includes the choice of quality measures, depends on the objective of the Digital Twin and the input/output data available (Muñoz et al., 2022). For the proposed use case of a Digital Twin it is essential that its virtual part is a reliable reflection of the real system, where anomalous outputs of the digital twin, or a discrepancy between the real and simulated outputs can be used to identify the onset of a failure in the system.

In this case the virtual part of the DT can be validated by feeding the same input data from the physical system into the model and using predefined criteria to assess how well the simulation replicates the real system (Hua et al., 2022). This can be done by quantifying the errors between the simulated and real outputs (Phillips & Kenley, 2024).

Once validated and deployed, a Digital Twin can use operational data to perform the same comparison to produce insights into the physical asset's state and evaluate the need for any action.

### 3.2.1. Absolute percent error

The Absolute Percentage Error (APE) is a straightforward calculation of the percentage deviation between the simulated/calculated value ($c$) and the true value (real data from sensors ($a$) at a particular timestamp. Expression for the metric is demonstrated in Equation 1 (Armstrong & Collopy, 1992):

$$APE = \left|\frac{a - c}{a}\right| \times 100\% \tag{1}$$

APE calculates the deviation at a particular time, meaning that previous calculated values do not affect subsequent ones, resulting in an independent metric. By definition, outliers and real values that are equal to or close to zero may lead to invalid APE values.

### 3.2.2. Mean absolute error (MAE) and mean squared error (MSE)

The Mean Absolute Error (MAE) is calculated as the sum of the absolute error divided by the number of samples, $n$. Equation 2 represents the expression for MAE:

$$MAE = \frac{1}{n}\sum\nolimits_{t=1}^{n} |a_t - c_t| \tag{2}$$

MAE is not sensitive to outliers, which makes it an appropriate choice for cases where occasional extreme values are expected. However, if a large number of outliers is present, the true quality of the model may be obscured. Additionally, MSE is not dimensionless and is not scaled, so interpretation of the result requires knowledge of the data distribution (Chicco et al., 2021).

On the other hand, by definition, the Mean Squared Error (MSE) amplifies outliers and extreme model predictions by squaring the difference between real and simulated values, making it suitable for applications where such occurrences need to be detected. MSE is calculated as follows (Equation 3) (Chicco et al., 2021):

$$MSE = \frac{1}{n}\sum\nolimits_{t=1}^{n} |a_t - c_t|^2 \tag{3}$$

Akin to MAE, MSE may be challenging to interpret due to its dimensional and squared outputs. Despite the mentioned drawbacks, there is evidence MAE, as well as MSE and its variations are used for DT validation. MSE has been used as a distance measure between virtual and real output data as part of a real-time model validation and updating framework with the aim to qualify the model as a Digital Twin (Emmert-Streib, 2023). MSE was also used to compare computed and real wind speeds to support the use of a DT as a virtual sensor for wind turbine generators (Ibrahim et al., 2023).

### 3.2.3. *Mean absolute percentage error (MAPE)*

The Mean Absolute Percentage Error (MAPE) is equivalent to the average of the absolute percentage errors (Vivas et al., 2020) and is calculated as shown in Equation 4 (Prayudani et al., 2019).

$$MAPE = \frac{1}{n}\sum\nolimits_{t=1}^{n} \left|\frac{a_t - c_t}{a_t}\right| \times 100\% = \frac{1}{n}\sum\nolimits_{t=1}^{n} APE_t \tag{4}$$

MAPE is widely used for assessing model quality as it produces readily understandable results expressed in percentages. However, by definition, MAPE is sensitive to outliers, which can magnify its value and obscure the true performance of the model (Tayman & Swanson, 1999).

### 3.2.4. *Symmetric mean absolute percentage error (SMAPE)*

The above are well-known metrics that are used to calculate the deviation between data for model fitting purposes. Given some of the drawbacks outlined, it can be argued that it is difficult to ascertain the performance of a model against these metrics without additional information. For example, MAE = 0 indicates a model that perfectly replicates the true output, while the upper bound is infinity. Therefore, Chicco et al. (2021) argue that MAE = 10 is not informative enough by itself to determine the quality of the model, as the worst possible outcome is unknown. To make that judgement, a predefined maximum acceptable MAE, or the distribution of true output values is needed. The same is true for MAPE and MSE (Chicco et al., 2021; Tayman & Swanson, 1999). The Symmetric Mean Absolute Percentage Error (SMAPE) eliminates this issue by forcing the values into [0, 200]% range, where 0% means perfect fit of the model to real data, and 200% – worst possible fit.

The authors point out that the definition of SMAPE is not consistent across literature and settle on the following expression (Equation 5) referring to several foundational works where the metric was introduced (Chicco et al., 2021):

$$SMAPE = \frac{1}{n}\sum\nolimits_{t=1}^{n} \frac{|c_t - a_t|}{(|c_t| + |a_t|)/2} \times 100\% \tag{5}$$

# 4. Results

## 4.1. Digital twin

Figure 4 shows the concept for the complete Digital Twin of the train braking system developed in this study. It is largely based on the STEM model (Figure 2) with some modifications tailored to the specific use case in this research.

Firstly, the *Observation* construct was placed inside the Virtual Space. As mentioned before, the nature of the *Observation* depends on the physical system and phenomena being monitored, the purpose of the

DT in terms of the product lifecyle stage, as well as the type of observer. In this study, the purpose of the DT is to monitor pressure in the pneumatic Feed Line of the train to assess the health of the braking system. The pressure predicted by the Digital Twin via simulation (*Physical Effects* and *Physical Phenomena* in Figure 4) represents normal behavior of the system. Therefore, the DT detects the onset of failure (*Observation*) in the system by detecting discrepancies (*Change of State*) between real and simulated pressure automatically and in real-time or near real-time. The comparison does not require input from a human and can be implemented fully virtually.

Secondly, in this case, the *Change of State* block requires sensor data from the Real Space to make the comparison. This is represented by the arrow connecting *External Inputs* and *Change of State* blocks in Figure 4. In this case, it could be seen that in terms of information, only the pressure values in the Feed Line are transferred to the *Change of State*.

Thirdly, a new type of information is introduced – the information bundle, represented by the double line in Figure 4. It means that the transmitted information consists of multiple signals. In this case, the whole bundle of information from the required sensors is transmitted from the Real Space to the Virtual Space, as well as from the sensors to the train's internal monitoring systems.

The components of the Virtual Space, including simulation of the Feed Line pressure (*Physical Effects* and *Physical Phenomena*), as well as the *Change of State* and *Observation* logic, were implemented in Simcenter Amesim. The Amesim model is a mathematical representation of the system functionality described in section 3.1. A detailed description of the model is outside the scope of this paper.

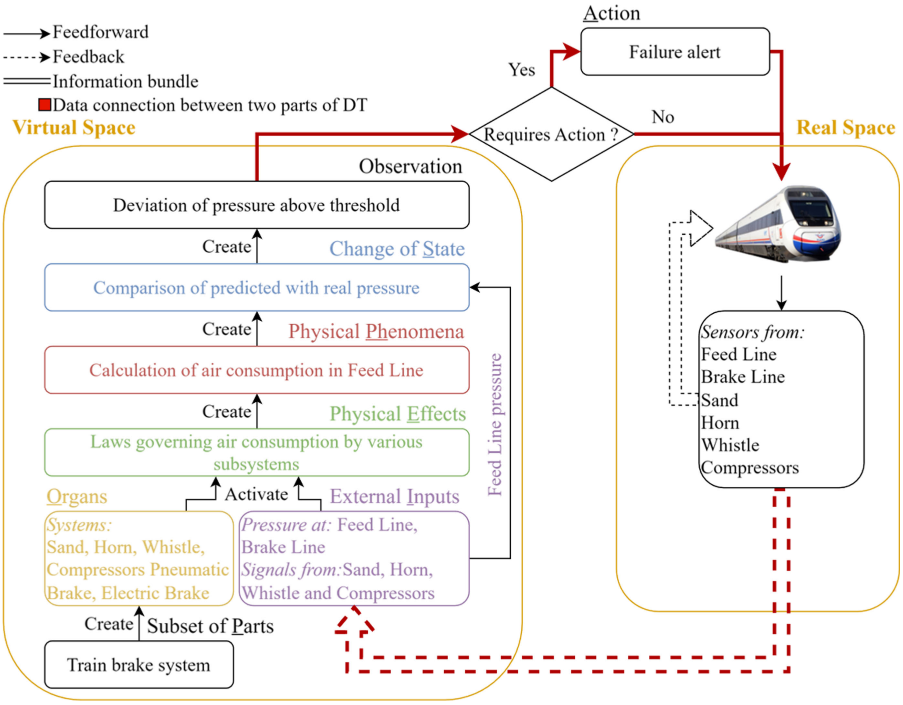
Feedforward
Feedback
Information bundle
🟥 Data connection between two parts of DT
Virtual Space
Action
Failure alert
Yes
Requires Action ?
No
Real Space
Observation
Deviation of pressure above threshold
Create
Change of State
Comparison of predicted with real pressure
Create
Physical Phenomena
Calculation of air consumption in Feed Line
Create
Physical Effects
Laws governing air consumption by various subsystems
Organs
Activate
External Inputs
Feed Line pressure
Systems: Sand, Horn, Whistle, Compressors Pneumatic Brake, Electric Brake
Pressure at: Feed Line, Brake Line
Signals from: Sand, Horn, Whistle and Compressors
Create
Subset of Parts
Train brake system
Sensors from:
Feed Line
Brake Line
Sand
Horn
Whistle
Compressors

Figure 4. Digital Twin of the train braking system

## 4.2. Validation

Successful implementation of functional and useful Digital Twins depends on their trustworthiness, which is supported by verification and validation. Validation tests the model's ability to replicate the real asset's behavior in relation to its intended purpose. It is an objective procedure that uses predefined criteria to classify a solution as validated. In this section, several quantitative measures, APE, MAE, MSE, MAPE and SMAPE, presented earlier, are applied to compare measured sensor data with the outputs of the DT described in the previous section. For this purpose, real input data, as seen in the *External Inputs* block in Figure 4, is used to run the simulation and produce virtual outputs.

Figure 5(a) shows a 4000-second long snapshot of a plot of FL pressure measured on the real train alongside FL pressure data predicted in the virtual part of the DT. Figure 5(b) shows a plot of APE in the time domain. The results of calculations of quality metrics on this data is shown in Table 1. The results are rounded to 2 decimal places.

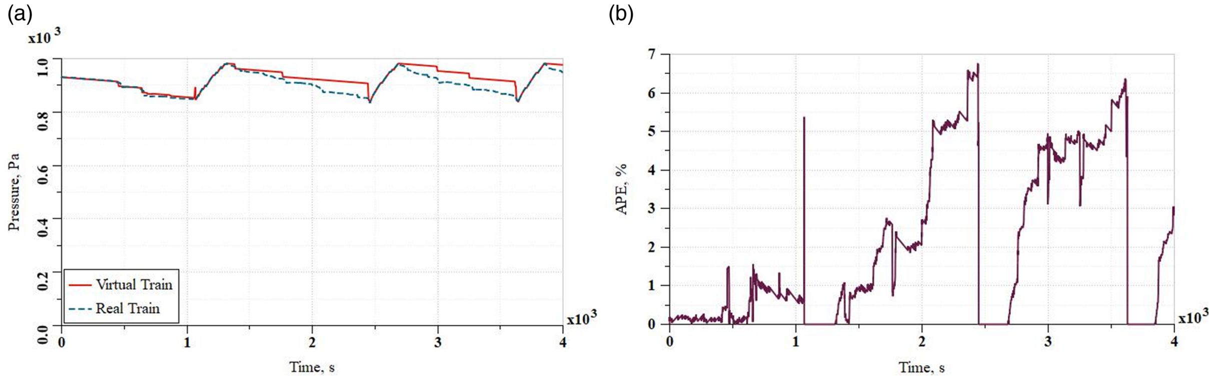

**Figure 5. (a) Feed Line pressure, (b) APE in the time domain**

**Table 1. Final result from quality metrics**

| Metric | Units | Value range | Result | Description |
|---|---|---|---|---|
| APE | dimensionless expressed in % | [0, + ∞) | 6.77 | Maximum instantaneous deviation between real and predicted data at the certain timestamp (t = 2441 sec.) |
| MAE | Pa | [0, +∞) | 17.92 | Average deviation between real and predicted data over the dataset |
| MSE | Pa$^2$ | [0, +∞) | 652.91 | Average squared deviation between real and predicted data over the dataset |
| MAPE | dimensionless expressed in % | [0, +∞) | 2.00 | Average deviation between real and predicted data over the dataset, expressed in percentages |
| SMAPE | dimensionless expressed in % | [0, 200] | 1.96 | Adjusted MAPE |

Each APE value on the graph represents the error magnitude at each timestamp. In other words, the metric does not take into account previous error values. Therefore, the last value of APE does not represent the overall quality of the DT. The graphical presentation of APE lets one identify points where the greatest errors occurred. These points occur just before the 2500-second mark and just after the 3500-second mark. These findings are consistent with Figure 5, which also shows that the largest deviations occur at these timestamps.

From Table 1, the final MAE value is equal to 17.92, which represents the average deviation of simulated outputs from real data. This is a dimensional value expressed in Pa. This result can be interpreted knowing the distribution of possible FL pressure values. Figure 5 shows that true FL pressure falls in the range between 800 and 1000 Pa, which means MAE suggests that the deviation is small. Meanwhile, MAPE is equal to 2.00%, indicating high accuracy. This is consistent with the result for MAE. Since the errors are squared in the MSE calculation, the magnitude of the resulting value is significantly higher. This makes it challenging to interpret the quality of the model from MSE, or correlate it to the other metrics. SMAPE shows consistency with MAPE in magnitude with the final value equal to just under 2%. The upper bound for SMAPE, i.e. the worst possible value, is 200%, which suggests the model is accurate.

From the above, MAPE and SMAPE are the most easily understandable measures of the overall model performance over the test dataset due to their dimensionless nature. MAPE results in a more conservative estimate of the model quality, as it is higher than SMAPE. Consequently, it is recommended to use MAPE to define validation criteria.

## 4.3. Decision-making

Metrics of the deviation between simulated and real sensor data are calculated in the *Change of State* block, as shown in Figure 4. Afterwards, the DT assesses the state of the physical asset by comparing the resulting metrics to their predefined thresholds and communicate this information as part of the decision-making and data exchange framework, following the STEM methodology, as outlined in previous sections. Real data containing a malfunction is not available. Therefore, for this purpose, virtual data (Figure 5(a)) was artificially amplified to create a new normal behavior benchmark. As a result, an artificial discrepancy between virtual and real data is created, which emulates a Feed Line pipe leak, as can be seen in Figure 6(a) within APE together with threshold Figure 6(b).

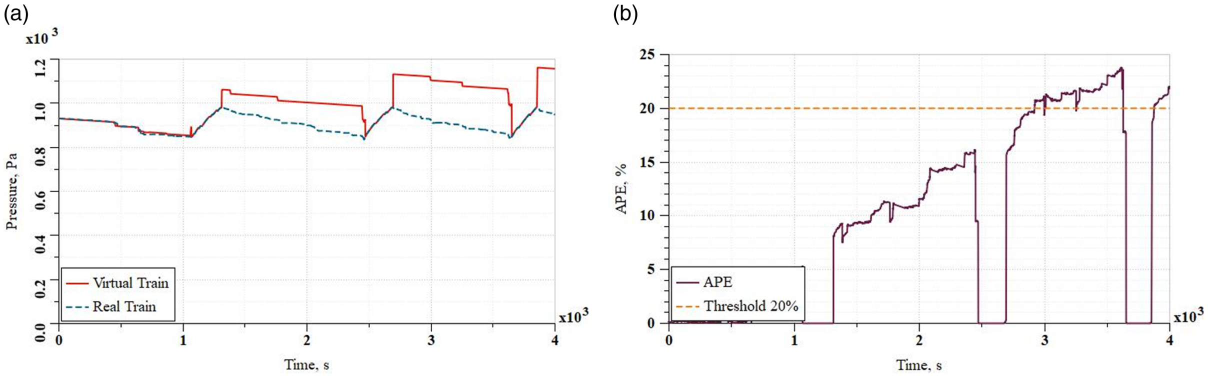

**Figure 6. (a) Physical phenomena: feed line pressure with leaks, (b) Observation: APE exceeds threshold just before 3000-second mark**

According to the proposed STEM Digital Twin schema (Figure 4), the DT compares real and simulated Feed Line pressure by quantifying errors (*Change of State*) between them and checking against thresholds to decide whether the system is behaving normally (*Observation*). This type of analysis brings the most value to users of the DT when operating in real-time or near real-time. To achieve this, the DT must be capable of identifying increased deviations instantaneously. While metrics that aggregate the error over a period of time provide a more comprehensive view of the model's overall quality, the Absolute Percentage Error (APE) is the most appropriate method for measuring instantaneous deviations without obscuring the true state of the real asset by averaging the errors. This allows the DT to provide timely insights to users, who act to investigate or eliminate the issue in the physical system.

The calculated *Change of State*, i.e. APE result, and an example threshold of 20% that is used to make an *Observation* are shown in Figure 6(b). Consequently, the *Observation* is that APE exceeds the threshold after some time. Therefore, a conclusion can be made about the onset of a failure in the physical system that causes an increased deviation between virtual and real data. This information is then fed forward to the physical world to alert users closing the information loop in the DT.

As a result, the proposed strategy implements both classic Digital Twin methodology and the extended STEM model to facilitate bi-directional information exchanges and decision-making. It enhances these processes with objective quantitative benchmarks, laying the foundation for standardization for applications across industries

## 5. Discussion

Having a validated Digital Twin is equally important for developers of DTs and their virtual parts, as well as for customers of the final DT product. However, the criteria for validation may differ. At a glance, any result out of APE, MAE, MAPE, or SMAPE (Table 1) on its own suggests that the DT captures the behavior of the real braking system very well, thus validating the solution. However, a formal validation procedure requires a clear benchmark for model performance. For example, a customer requirement might specify that the final DT product is validated if the Absolute Percentage Error (APE) does not exceed 20% on the test dataset. Based on the results presented in Table 1 and Figure 5(b), the Digital Twin developed in this study meets this requirement as the maximum APE is less than 7%. Understandably, the choice of the validation dataset is an important consideration. A larger dataset was also studied as part of this investigation. The findings suggest that quantifying metrics as part of a validation strategy can be a valuable tool for identifying aspects – such as data quality issues, or system behaviors – that may have been overlooked initially but affect model fidelity. In other words, this tool can lead to better understanding of the system, data collection methods, and specific areas where improvements could be made to the DT.

In the presented analysis, the metrics were examined in combination. It was demonstrated that the results for these five metrics can be mutually correlated, providing a form of cross-validation. This led to the recommendation that the Mean Absolute Percentage Error (MAPE) is the most suitable validation metric. This recommendation was a result of evidence from several calculated metrics. So, in reality, the validation did not rely on a single arbitrarily chosen measure of model performance. Therefore, it can be argued that using a combination of metrics to assess the performance of a Digital Twin allowed for a more informed and well-rounded decision on the validation procedure.

APE was proposed as the most appropriate choice of metric for the Digital Twin to facilitate decision-making, as it provides an estimate of the physical system's immediate state. This is a relevant argument for the current use case of a railway braking system failure, such as a feed line pipe leak, where an instantaneous decrease in feed line pressure is indicative of this type of malfunction. This type of decision-making can apply to other engineering systems where sudden deviations point to a malfunction. This reasoning does not always apply. For example, instant drops in the train cabin temperature may be a normal response to the opening and closing of doors during train operation. In that case, MAPE may be a more appropriate metric as it can be used to indicate whether the average temperature deviation throughout the day exceeded a specified threshold to detect a malfunctioning climate control system. However, when using metrics like MAPE, which rely on averages, it is important to select a representative time period that adequately captures the onset of failure. This example highlights that the selection of a quantitative metric for decision-making within a Digital Twin depends on the particular use case. While the proposed methodologies for validation and decision-making are not limited to the train braking system or the railway industry, the use of quantitative metrics relies on time series data.

## 6. Conclusion

This paper presents an analysis of several quantitative metrics as quality measures for validation of virtual parts of Digital Twins, as well as implementation of them in a Digital Twin's feedforward decision-making. The DT of a railway braking system was developed based on the modified STEM DT concept. The STEM methodology, which is an extension of the SAPPhIRE model, adds value to the design of the validation procedure and the decision-making framework, as the resulting Digital Twin has a clear structure with defined cause-effect relationships. It was found that an analysis of multiple metrics to assess the performance of a DT can help developers in creating a more comprehensive validation strategy, laying the foundation for standardization in the area of DT verification and validation and contributing to a more robust final DT solution. For the DT developed as part of this research, the described approach led to the selection of the Mean Absolute Percentage Error (MAPE) as the most appropriate quantitative metric as an objective validation criterion.

This work presented an analysis of these metrics with the aim of integration into the DT decision-making framework. It was demonstrated that metric selection depends on aspects of the Digital Twin use case. For example, the physical phenomena and systems being monitored, as well as the intended outcome of the DT's decision-making, dictate whether instantaneous or average divergence is more indicative of the physical asset's condition. Additionally, customers of DT products may exhibit a preference towards percentage-based indicators, as they are more intuitive, or provide acceptable ranges of normal behavior based on their knowledge of the asset. As a result, the Absolute Percentage Error (APE) was selected as a threshold to support the decision-making of the braking system Digital Twin.

The outlined approach to designing the decision-making framework of a DT could be extended to the development of Predictive Maintenance strategies. For instance, results presented in Figure 6(b) indicate an upward trend in APE, suggesting that a failure could be anticipated before the APE intersects with the threshold. Therefore, implementation of strategies based on the STEM concept can facilitate an informed approach towards Predictive Maintenance design. Further research could explore novel ways for incorporating these metrics into more advanced decision-making. For example, a more well-rounded assessment of the physical system's health could involve monitoring whether a metric has been approaching the threshold for an extended period, as this may indicate a pre-failure condition. In this case, given sufficient historic data, advanced techniques, such as Machine Learning, may be integrated to determine the nature of the failure based on the behavior of the metric over time.

## References

Armstrong, J. S., & Collopy, F. (1992). Error measures for generalizing about forecasting methods: Empirical comparisons. *International Journal of Forecasting*, 8(1), 69–80. https://doi.org/10.1016/0169-2070(92)90008-W

Bhattacharya, K., & Chakrabarti, A. (2023). Application of SAPPhIRE Model of Causality in the Design of Product-Service Systems. 527–539. https://doi.org/10.1007/978-981-99-0428-0_43

Bitencourt, J., Osho, J., Harris, G., Purdy, G., & Moreira, A.C. (2023). Building Trust in Digital Twin through Verification and Validation. *IISE Annual Conference & Expo 2023*.

Boschert, S., & Rosen, R. (2016). Digital Twin—The Simulation Aspect. In Hehenberger P. & Bradley D. (Eds.), *Mechatronic Futures: Challenges and Solutions for Mechatronic Systems and their Designers* (pp. 59–74). Springer International Publishing. https://doi.org/10.1007/978-3-319-32156-1_5

Chakrabarti, A., Sarkar, P., Leelavathamma, B., & Nataraju, B.S. (2004). A functional representation for aiding biomimetic and artificial inspiration of new ideas. AI EDAM, 19(2), 113–132. https://doi.org/10.1017/S0890060405050109

Chicco, D., Warrens, M.J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, 7, 1–24. https://doi.org/10.7717/PEERJ-CS.623/SUPP-1

Emmert-Streib, F. (2023). Defining a Digital Twin: A Data Science-Based Unification. *Machine Learning and Knowledge & Extraction*. https://doi.org/10.3390/make5030054

Ershenko, D., Sadeghzadeh, S., Fortin, C., & Panayi, A. (2024). On the integration of the SAPPhIRE model in the Digital Twin development process: A train braking system use case. *PLM* 2024

Grieves, M. (2015). Digital Twin: Manufacturing Excellence through Virtual Factory Replication

Grieves, M. (2023). Digital Twin Certified: Employing Virtual Testing of Digital Twins in Manufacturing to Ensure Quality Products. *Machines*, 11(8), Article 8. https://doi.org/10.3390/machines11080808

Grieves, M., & Vickers, J. (2017). Digital Twin: Mitigating Unpredictable, Undesirable Emergent Behavior in Complex Systems. *Transdisciplinary Perspectives on Complex Systems: New Findings and Approaches*, 85–113. https://doi.org/10.1007/978-3-319-38756-7_4

Hua, E. Y., Sanja, L.-M., & Francis, D. (2022). Validation of Digital Twins: Challenges and Opportunities. *2022 Winter Simulation Conference*

Ibrahim, M., Rassõlkin, A., Vaimann, T., Kallaste, A., Zakis, J., Hyunh, V.K., & Pomarnacki, R. (2023). Digital Twin as a Virtual Sensor for Wind Turbine Applications. 11, 6246. https://doi.org/10.3390/en16176246

Li, L., Aslam, S., Wileman, A., & Perinpanayagam, S. (2022). Digital Twin in Aerospace Industry: A Gentle Introduction. *IEEE Access*, 10, 9543–9562. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3136458

Ma, S., Flanigan, K.A., & Bergés, M. (2024). State-of-the-art review and synthesis: A requirement-based roadmap for standardized predictive maintenance automation using digital twin technologies. *Advanced Engineering Informatics*, 62, 102800. https://doi.org/10.1016/j.aei.2024.102800

McSorley, G., Fortin, C., & Huet, G. (2014). Modified SAPPhIRE model as a framework for Product Lifecycle Management. *DS 77: Proceedings of the DESIGN 2014 13th International Design Conference*, 1843–1852.

Muñoz, P., Wimmer, M., Troya, J., & Vallecillo, A. (2022). Using trace alignments for measuring the similarity between a physical and its digital twin. *Proceedings of the 25th International Conference on Model Driven Engineering Languages and Systems: Companion Proceedings*, 503–510. https://doi.org/10.1145/3550356.3563135

Phillips, I., & Kenley, C.R. (2024, June 7). Validation Framework of a Digital Twin: A System Identification Approach. *34th Annual INCOSE international symposium, Dublin, Ireland.*

Prayudani, S., Hizriadi, A., Lase, Y.Y., & Fatmi, Y. (2019). Analysis Accuracy Of Forecasting Measurement Technique On Random K-Nearest Neighbor (RKNN) Using MAPE And MSE. *Journal of Physics: Conference Series*, 1361(1), 012089. https://doi.org/10.1088/1742-6596/1361/1/012089

Pylianidis, C., Osinga, S., & Athanasiadis, I.N. (2021). Introducing digital twins to agriculture. *Computers and Electronics in Agriculture*, 184, 105942. https://doi.org/10.1016/j.compag.2020.105942

Srinivasan, V., & Chakrabarti, A. (2009). SAPPhIRE - an approach to analysis and synthesis. *ICED*, 2, 417–428

Tayman, J., & Swanson, D.A. (1999). On the validity of MAPE as a measure of population forecast accuracy. *Population Research and Policy Review*, 18(4), 299–322. https://doi.org/10.1023/A:1006166418051/METRICS

Villalonga, A., Negri, E., Fumagalli, L., Macchi, M., Castaño, F., & Haber, R. (2020). Local Decision Making based on Distributed Digital Twin Framework. *IFAC-Papers On Line*, 53(2), 10568–10573. https://doi.org/10.1016/j.ifacol.2020.12.2806

Vivas, E., Allende-Cid, H., & Salas, R. (2020). A Systematic Review of Statistical and Machine Learning Methods for Electrical Power Forecasting with Reported MAPE Score. *Entropy 2020*, Vol. 22, Page 1412, 22(12), 1412. https://doi.org/10.3390/E22121412

West, S., Stoll, O., Meierhofer, J., & Züst, S. (2021). Digital Twin Providing New Opportunities for Value Co-Creation through Supporting Decision-Making. 11, 3750. https://doi.org/10.3390/app11093750