Challenges of Foundation Models for Physics-Based Domains: Insights from the TURING Project

Matic Medvešek†
Rudolfovo - Science and Technology Centre Novo Mesto
Novo mesto, Slovenia
matic.medvesek@rudolfovo.eu

Jelena Topić Božič
Rudolfovo - Science and Technology Centre Novo Mesto
Novo mesto, Slovenia
jelena.topic.bozic@rudolfovo.eu

Simon Muhič
Rudolfovo - Science and Technology Centre Novo Mesto
Novo mesto, Slovenia
simon.muhic@rudolfovo.eu

Abstract

Foundation models (FMs) are increasingly applied to scientific computing, offering opportunities to accelerate computationally demanding simulations and support scientific decision-making. However, physics-based applications impose requirements extending beyond predictive accuracy, including physical consistency, generalisation, uncertainty quantification and compatibility with established numerical workflows. This paper discusses these challenges based on insights from the TURING project, focusing on two use cases: reactor modelling at NRG PALLAS and regional weather forecasting at Météo-France. Despite substantial differences in physical scale, data availability and operational objectives, both applications expose common limitations of current foundation-model approaches. The analysis indicates that FMs are most promising as physics-aware components of hybrid computational workflows rather than direct replacements for established numerical models.

Keywords

foundation models, scientific artificial intelligence, physics-informed learning, simulation, uncertainty

1 Introduction

Foundation models (FMs) have demonstrated considerable transferability across natural language, computer vision and increasingly scientific applications. Recent advances in data-driven weather modelling, including GraphCast [1] and Pangu-Weather [2], together with foundation-model approaches such as ClimaX [3], AtmoRep [4] and Aurora [5] demonstrate the potential of large pretrained architectures for representing complex physical systems.

Physics-based applications, however, differ fundamentally from conventional artificial intelligence (AI) tasks. Their behaviour is constrained by governing equations, conservation principles and boundary conditions, while numerical predictions may support safety- or mission-critical decisions [6].

The Trustworthy Unified Robust Intelligent Generative Systems (TURING) project investigates these requirements across nuclear engineering, high-energy physics and meteorology. This work focuses on two contrasting cases: NRG PALLAS reactor modelling and Météo-France regional weather forecasting. NRG PALLAS represents a safety-critical, comparatively data-constrained engineering system, whereas weather forecasting represents a data-rich, high-dimensional spatio-temporal problem. Their comparison enables common challenges of FMs in physics-based domains to be identified independently of the physical system.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, 8–9 October 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
http://doi.org/10.70314/is.2026.skui.58

2 Representative TURING Use Cases

2.1 NRG PALLAS: Safety-Critical Reactor Modelling

The NRG PALLAS use case considers transient thermal-hydraulic prediction and steady-state reactor core modelling. The former concerns classification and prediction of abnormal reactor behaviour, including rapid reactor shutdown (SCRAM) events, while the latter maps reactor configurations to neutronic quantities including neutron flux, power distribution and the effective multiplication factor K_{eff}.

The modelling workflows rely strongly on data generated using established numerical simulation methods. Realistic full-core calculations may require approximately one hour of high-performance computing (HPC) resources, making individual samples considerably more expensive than typical machine-learning data. At the same time, the safety-critical nature of reactor modelling requires reliable uncertainty estimation and identification of conditions outside the training distribution. Physics-informed approaches [6], conformal prediction [7], Monte Carlo dropout (MC-dropout) [8] and inference-time out-of-distribution (OOD) diagnostics [9] therefore represent relevant mechanisms for model trustworthiness.

2.2 Météo-France: Regional Weather Forecasting

The Météo-France use case targets short-range regional weather forecasting over Western Europe at kilometre-scale resolution, with the Application of Research to Operations at Mesoscale (AROME) model providing the established high-resolution numerical reference [10]. Global data-driven models including GraphCast [1] and Pangu-Weather [2] have demonstrated that learned atmospheric representations can provide competitive medium-range forecasts at substantially lower inference cost than conventional numerical weather prediction. These models demonstrate the ability of learned representations to capture

large-scale atmospheric evolution, but their application to regional forecasting introduces additional requirements. Kilometre-scale prediction must preserve local spatial structures and processes that may not be sufficiently represented at the resolution of global training data. These developments provide an important basis for more general pretrained Earth-system models, including ClimaX [3], AtmoRep [4] and Aurora [5]. However, transferring data-driven models to convection-permitting regional forecasting while maintaining small-scale physical structures and calibrated uncertainty remains challenging [11], [12], [13]. Probabilistic forecasting approaches such as GenCast [14] further demonstrate the increasing importance of uncertainty-aware prediction.

The two use cases place very different demands on a foundation model. Table 1 summarises these differences and, more importantly, shows where the same methodological problems emerge despite the different physical systems.

Table 1: Comparison of the two representative TURING use cases

Characteristic	NRG PALLAS	Météo-France
Domain	Reactor neutronics and thermal-hydraulics	Regional weather forecasting
Main task	Core modelling and transient/SCRAM prediction	Kilometre-scale short-range forecasting
Data	Simulated core data and thermal-hydraulic variables	Large, high-dimensional atmospheric datasets
Reference model	MGRAC/OSCAR, TRACE	AROME
Data generation	Computationally expensive; relatively limited	Large-scale and data-rich
Physical challenge	Coupled reactor physics and transients	Chaotic, multiscale atmospheric dynamics
Generalisation	New configurations and operating conditions	Global-to-regional transfer
Trustworthiness	OOD detection and safety-related uncertainty	Forecast uncertainty and local extremes
Potential FM role	Fast surrogate and state prediction	Fast regional forecasting

3 Common Challenges

The nuclear reactor use case is strongly constrained by the cost and availability of high-fidelity data, whereas weather forecasting has access to much larger datasets but must represent atmospheric dynamics over a wide range of spatial and temporal scales. Nevertheless, the practical questions are similar: does the model preserve the relevant physics, does it remain reliable outside familiar conditions, can its uncertainty be quantified, and when should the numerical model still be used? These questions form the basis of the following discussion.

3.1 Physical Consistency

Unlike conventional numerical solvers, FMs learn statistical representations without necessarily satisfying the governing equations. A prediction may therefore exhibit low statistical error while violating conservation laws or other physical constraints. Physics-informed approaches have demonstrated the potential of incorporating physical constraints into surrogate models for nuclear applications [6], while physics-aware loss functions have been investigated for preserving spectral and spatial characteristics in weather prediction [11].

Physical consistency should be considered during both training and validation. Physical knowledge can be incorporated through training data, loss functions, architectural constraints or coupling with numerical solvers. Figure 1 illustrates one such approach, in which physical constraints are incorporated into the training loss of a physics-informed neural network. Training data may implicitly capture physical relationships within the represented operating domain but do not ensure their preservation during extrapolation to unseen physical regimes. Explicit constraints can reduce this risk by penalising physical inconsistencies. For safety-critical applications, physical consistency should therefore be evaluated independently from conventional prediction error, particularly under transient and off-design conditions [6].

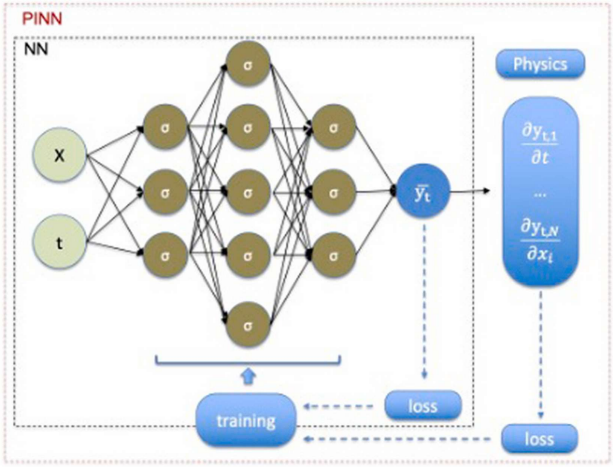

Figure 1: Physics-informed neural-network framework incorporating physical constraints into model training. Adapted from [6] under the CC BY 4.0 license.

3.2 Generalisation and Transferability

Transferability is a defining objective of foundation models but remains challenging in physical systems, where changes in geometry, operating conditions, boundary conditions or spatial scale can modify the governing behaviour. Current nuclear models remain largely reactor-specific, while global weather models require substantial adaptation for kilometre-scale regional forecasting [12], [13]. Scientific generalisation should therefore be evaluated across unseen physical regimes, rather than solely through random train-test subdivision.

Evaluation should include structured distribution shifts involving unseen geometries, boundary conditions, parameter ranges or spatial resolutions. Whether FMs can transfer learned representations between related physical systems remains an important open question.

3.3 Uncertainty and Trustworthiness

Scientific predictions require information about when a model can reasonably be trusted. For nuclear reactor use case, uncertainty is directly associated with safety-related decisions, motivating approaches such as conformal prediction [7], MC-dropout [8] and OOD detection [9]. In weather forecasting, uncertainty arises additionally from chaotic atmospheric dynamics and is conventionally represented through ensembles. Probabilistic models such as GenCast [14] illustrate how this requirement is increasingly being incorporated into data-driven forecasting.

MC-dropout approximates epistemic uncertainty through repeated stochastic inference [8], conformal prediction provides statistically calibrated prediction intervals under appropriate assumptions [7], and OOD diagnostics assess whether new inputs differ substantially from the training distribution [9]. Since uncertainty and OOD behaviour are not equivalent, their combination can provide both predictive uncertainty information and an indication of whether a case lies within the demonstrated applicability domain. Such information can determine when high-fidelity numerical verification is required.

3.4 Integration with Physics-Based Simulation

The analysed requirements suggest that FMs are more appropriately considered as complements to established numerical simulations than as direct replacements. For reactor states represented by the training, data could be evaluated rapidly by the learned model, while uncertain or OOD cases could be passed to the high-fidelity solver. A similar principle applies to regional weather modelling, where numerical models and observations remain important references for assessing forecast quality. Such a hybrid approach concentrates expensive numerical calculations on cases where they provide the greatest additional information. The results of these calculations could subsequently be incorporated into model development, gradually extending the range of represented physical conditions. This interaction creates an adaptive workflow in which the FM provides computationally efficient predictions within its demonstrated applicability domain, while the numerical model remains available for verification of uncertain or insufficiently represented cases. Such integration is particularly relevant when high-fidelity data are computationally expensive to generate, as in the PALLAS use case, and provides a practical mechanism for combining the computational efficiency of learned models with the reliability of established simulation methods.

4 Insights and Future Directions

Evaluation of scientific FMs should extend beyond conventional accuracy metrics to include application-specific physical quantities, uncertainty calibration and structured extrapolation tests [7], [9]. For the nuclear reactor task, a low aggregate prediction error would not be sufficient if the surrogate failed to preserve safety-relevant thermal-hydraulic trends or identify reactor states outside its validated range. For regional forecasting, a comparable issue arises when a model preserves large-scale fields but underrepresents local extremes or poorly calibrates forecast uncertainty. Weather forecasting may therefore emphasise spatial characteristics, extremes and probabilistic skill [2], [14], whereas reactor applications require particular attention to safety-relevant quantities, transient behaviour and operational limits [6].

Scientific FMs should also have a defined applicability domain analogous to the documented assumptions and validation ranges of conventional numerical models. This domain may describe the physical regimes, parameter ranges, spatial and temporal resolutions and OOD conditions represented during model development [9]. It would enable users to distinguish interpolation within a demonstrated operating envelope from extrapolation into insufficiently represented regimes. Predictions near or outside this envelope could trigger uncertainty analysis or high-fidelity verification [7], linking established verification and validation practices with emerging FM methodologies.

The computational benefit of scientific FMs also depends on model reuse. Although inference may be substantially faster than high-fidelity simulation, generating training data and pretraining large models can itself be expensive. This balance differs between the two cases. Weather forecasting can exploit large datasets and repeated operational forecasts, whereas reactor training samples may require expensive high-fidelity calculations [2], [6], [14]. For PALLAS, the value of pretraining therefore depends strongly on whether learned representations can be reused across reactor configurations, operating states or related tasks. A model requiring extensive retraining for each configuration offers considerably less advantage over a problem-specific surrogate.

5 Conclusions

Two contrasting TURING applications reactor modelling at NRG PALLAS in the Netherlands and Météo-France regional weather forecasting were used to examine the challenges of foundation models in physics-based domains. Despite major differences in physical scale and data availability, four overarching requirements were synthesised from the comparison:

- physical consistency,
- generalisation and transferability,
- uncertainty and trustworthiness,
- integration with established simulation workflows.

The relative importance of these requirements depends on the application. Data-rich meteorological applications primarily expose challenges associated with scale transfer, representation of small-scale processes and probabilistic prediction, whereas reactor modelling places stronger emphasis on limited high-fidelity data, OOD behaviour and safety-related uncertainty. Nevertheless, scientific FMs should provide information about the applicability and reliability of their predictions in addition to the predictions themselves.

Overall, the near-term value of scientific FMs appears to lie primarily in augmenting rather than replacing physics-based simulation. Combining transferable representations with physical constraints, uncertainty quantification, OOD detection and domain-specific validation provides a promising pathway towards trustworthy AI-assisted scientific modelling.

Acknowledgments

The research was funded by the European Union's Horizon Europe research and innovation programme through the TURING Project. Call HORIZON-CL4-2024-HUMAN-03-02 – Explainable and Robust AI.

References

[1] R. Lam *et al.*, "Learning skillful medium-range global weather forecasting," *Science (1979).*, vol. 382, no. 6677, pp. 1416–1421, Dec. 2023, doi: 10.1126/science.adi2336.

[2] K. Bi, L. Xie, H. Zhang, X. Chen, X. Gu, and Q. Tian, "Accurate medium-range global weather forecasting with 3D neural networks," *Nature*, vol. 619, no. 7970, pp. 533–538, Jul. 2023, doi: 10.1038/s41586-023-06185-3.

[3] T. Nguyen, J. Brandstetter, A. Kapoor, J. K. Gupta, and A. Grover, "ClimaX: A foundation model for weather and climate," in *Proceedings of the 40th International Conference on Machine Learning*, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., in Proceedings of Machine Learning Research, vol. 202. PMLR, Aug. 2023, pp. 25904–25938. [Online]. Available: https://proceedings.mlr.press/v202/nguyen23a.html

[4] C. Lessig, I. Luise, B. Gong, M. Langguth, S. Stadler, and M. Schultz, *AtmoRep: A stochastic model of atmosphere dynamics using large scale representation learning*. 2023. doi: 10.48550/arXiv.2308.13280.

[5] C. Bodnar *et al.*, "A foundation model for the Earth system," *Nature*, vol. 641, no. 8065, pp. 1180–1187, 2025, doi: 10.1038/s41586-025-09005-y.

[6] F. Antonello, J. Buongiorno, and E. Zio, "Physics informed neural networks for surrogate modeling of accidental scenarios in nuclear power plants," *Nuclear Engineering and Technology*, vol. 55, no. 9, pp. 3409–3416, Sep. 2023, doi: 10.1016/j.net.2023.06.027.

[7] A. Angelopoulos and S. Bates, *A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification*. 2021. doi: 10.48550/arXiv.2107.07511.

[8] Y. Gal and Z. Ghahramani, "Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning," *Proceedings of The 33rd International Conference on Machine Learning*, Jun. 2016.

[9] H. Wang and Y. Li, "Bridging OOD detection and generalization: a graph-theoretic view," in *Proceedings of the 38th International Conference on Neural Information Processing Systems*, in NIPS '24. Red Hook, NY, USA: Curran Associates Inc., 2024.

[10] Y. Seity *et al.*, "The AROME-France Convective-Scale Operational Model," *Mon. Weather Rev.*, vol. 139, no. 3, pp. 976–991, 2011, doi: https://doi.org/10.1175/2010MWR3425.1.

[11] T. Dunstan *et al.*, "FastNet: Improving the Physical Consistency of Machine Learning Weather Prediction Models through Loss Function Design," *Artificial Intelligence for the Earth Systems*, vol. 5, no. 3, Jul. 2026, doi: 10.1175/AIES-D-25-0090.1.

[12] T. N. Nipen *et al.*, "Regional Data-Driven Weather Modeling with a Global Stretched Grid," *Artificial Intelligence for the Earth Systems*, vol. 5, no. 2, Apr. 2026, doi: 10.1175/AIES-D-25-0001.1.

[13] E. M. Nordhagen *et al.*, "High-Resolution Probabilistic Data-Driven Weather Modeling with a Stretched-Grid," *ArXiv*, vol. abs/2511.23043, 2025, [Online]. Available: https://api.semanticscholar.org/CorpusID:283438519

[14] I. Price *et al.*, "Probabilistic weather forecasting with machine learning," *Nature*, vol. 637, no. 8044, pp. 84–90, Jan. 2025, doi: 10.1038/s41586-024-08252-9.