# ISDA-Online

**July 10, 2025, from 15-17 UTC**

# "Machine Learning Data Assimilation"

**Organizers and Conveners:** ***Hristo Chipilski*** *(Florida State University, US),* ***Rossella Arcucci*** *(Imperial College London, UK),* ***Sibo Cheng*** *(CEREA, ENPC, Institut Polytechnique de Paris, France),* ***Tobias Necker*** *(ECMWF, Germany)*

*Learning from observations - Machine Learning (ML) or Data Assimilation (DA)? During the last few years, we have seen remarkable advancements in applying ML in the domain of DA. Combining ML and DA has already enabled the development of hybrid approaches that challenge long-standing pure DA systems. This event aims to explore recent developments in these directions. We welcome all abstract submissions that push the boundaries of machine learning applications in the context of data assimilation.*

**Program: (UTC)**

| | |
|---|---|
| **15:00 – 15:05** | **Welcome** |
| **15:05 – 15:30** | **AI-DOP: Learning a medium-range weather forecast directly from observations** |
| (20' + 5') | Boucher et al. |
| **15:30 – 15:50** | **Toward the development of coupled carbon and water cycle land data assimilation in the ECMWF Integrated Forecast System (IFS) by leveraging machine learning and new types of Earth observations** |
| (17' + 3') | Garrigues et al. |
| **15:50 – 15:55** | *time buffer* |
| **15:55 – 16:15** | **Bayesian inference for geophysical fluid dynamics using generative models** |
| (17' + 3') | Lobbe et al. |
| **16:15 – 16:35** | **Towards real-time prediction with autoencoders** |
| (17' + 3') | Özalp et al. |
| **16:35 – 16:55** | **A Multi-Fidelity Ensemble Kalman Filter with a machine learned surrogate model** |
| (17' + 3') | Van der Voort et al. |
| **16:55 – 17:00** | **Closing: Information on upcoming sessions** |

Please note:

- When you login to the session before 15:00 UTC, and everything is quiet, this is most likely because we muted the microphones.
- The times in UTC are approximate. In case of technical problems, we might have to change the order of the presentations.
- **Time Zones**: 15 – 17 UTC

  Europe: 04 – 06 pm BST (London) | 05 – 07 pm CEST (Berlin)
  Asia/Australia: 11 – 01 am CST (Shanghai) | 00 – 02 am JST (Tokyo) | 01 – 03 am AEDT (Sydney)
  Americas: 08 – 10 am PDT (San Fran.) | 09 – 11 am MDT (Denver) | 11 – 01 pm EDT (New York)

# AI-DOP: Learning a medium-range weather forecast directly from observations

Eulalie Boucher[1], Mihai Alexe[1], Peter Lean[1], Ewan Pinnington[1], Patrick Laloyaux[1], Anthony McNally[1], Simon Lang[1], Matthew Chantry[1], Chris Burrows[1], Marcin Chrust[1], Florian Pinault[1], Ethel Villeneuve[1], Niels Bormann[1], Sean Healy[1]

[1] ECMWF, Reading, UK / Bonn, Germany

In recent years, global data-driven numerical weather prediction (NWP) models have begun matching the performance of leading physics-based systems like ECMWF's IFS across key skill metrics. These AI models currently rely on weather reanalyses (e.g. ERA5) for training and require an analysis valid at the initial time of the forecast for initialization. These are obtained with, for instance, ECMWF's 4D-Var, assimilating over 20 million observations during each 12-hour cycle. While an undeniable success, 4D-Var is computationally expensive and requires tuning and specifications of complex backgrounds, observation error covariances, observation operators, tangent and adjoint model linearisations, variational bias corrections and forecast model error. An obvious question that arises is can machine learning offer an alternative? In response, ECMWF has been exploring an innovative data-driven method for medium-range forecasting, AI-Direct Observation Prediction (AI-DOP). This approach seeks to directly learn Earth System dynamics and processes by analyzing relationships among observational data (e.g., brightness temperatures) and geophysical variables like temperature and winds from weather stations. AI-DOP relies solely on historical time series of satellite and conventional observations, avoiding the use of climatological data or inputs derived from NWP (re)analysis. Early evaluations of AI-DOP forecasts indicate that the system can develop a consistent internal representation of the Earth System. The model effectively generalizes relationships between observed variables to regions without direct observational coverage—for instance, forecasts of upper-level winds in areas with sparse radiosonde or aircraft data align well with ERA5 reanalysis. We present an overview of AI-DOP, its current state of development, and the potential pathways to achieving a fully data-driven, end-to-end system capable of rivalling IFS for medium-range weather forecasting.

# Towards real-time prediction with autoencoders

Elise Özalp[1], Andrea Nóvoa[1], Luca Magri[1,2,3]

[1] Imperial College London
[2] The Alan Turing Institute
[3] Politecnico di Torino

Convolutional autoencoder recurrent neural networks have become a promising approach for forecasting chaotic and turbulent systems [1]. These models can effectively capture the stability properties of chaotic dynamics and forecast the short-term with high accuracy but might struggle with long-term autonomous simulations, for which the forecasts might become unstable [2, 3]. To address this, data assimilation techniques can be employed to combine model predictions with observations, correcting and stabilizing forecasts over extended time horizons. We propose a framework that integrates a hybrid convolutional autoencoder echo state network (CAE-ESN) model with the ensemble Kalman filter (EnKF) to achieve time-accurate, long-term predictions of chaotic systems. The CAE reduces high-dimensional system states to a lower-dimensional latent manifold, enabling efficient representation of the dynamics. The ESN autonomously predicts the evolution of the latent states, and sparse, noisy observations of the full state are used to correct the forecast by applying the EnKF. We demonstrate this approach on the chaotic Kuramoto-Sivashinsky equation and show that the EnKF augmented CAE-ESN framework achieves stable long-term predictions, maintaining time-accurate forecasts, whereas other approaches, such as the CAE-ESN alone, diverge after about 2 Lyapunov times. Further results will be presented for higher-dimensional turbulent systems, such as the Kolmogorov flow. This work opens up new opportunities for the long-term forecasting of chaotic and turbulent dynamics using autoencoding strategies.

[1] P. R. Vlachas, et al. Multiscale simulations of complex systems by learning their effective dynamics. Nature Machine Intelligence, 2022.
[2] A. Racca, et al. Predicting turbulent dynamics with the convolutional autoencoder echo state network. Journal of Fluid Mechanics, 2023.
[3] E. Özalp and L. Magri. Stability analysis of chaotic systems in latent spaces. Nonlinear Dynamics, 2024.

# Bayesian inference for geophysical fluid dynamics using generative models

Alexander Lobbe[1], Dan Crisan[1], Oana Lang[2]

[1] Imperial College London, UK
[2] Babeș-Bolyai University, Cluj-Napoca, Romania

Data assimilation plays a crucial role in numerical modeling, enabling the integration of real-world observations into mathematical models to enhance the accuracy and predictive capabilities of simulations. However, calibrating high-dimensional, nonlinear systems remains challenging.

This paper presents a novel calibration approach using diffusion generative models to produce synthetic data that align with observed numerical solutions of a stochastic partial differential equation. These samples enable efficient model reduction, assimilating data from a high-resolution rotating shallow water equation with 10^4 degrees of freedom into a reduced stochastic system with significantly fewer degrees of freedom.

The synthetic samples are integrated into a particle filtering method, enhanced with tempering and jittering, to handle complex, multimodal distributions. Our results demonstrate that generative models improve particle filter accuracy, offering a more computationally efficient solution for data assimilation and model calibration.

# Toward the development of coupled carbon and water cycle land data assimilation in the ECMWF Integrated Forecast System (IFS) by leveraging machine learning and new types of Earth observations

Sébastien Garrigues[1], Patricia de Rosnay[1], Peter Weston[1], David Fairbairn[1], Ewan Pinnington[1], Souhail Boussetta[1], Anna Agusti-Panareda[1], Jean-Christophe Calvet[2], Cédric Bacour[3], Richard Engelen[1]

[1] ECMWF, Reading, UK
[2] CNRM, Université de Toulouse, Météo-France, CNRS, France
[3] LSCE, Gif-Sur-Yvette, France

In this seminar, we present results on the assimilation at global scale of the normalized backscatter at 40° from ASCAT and SIF from TROPOMI, using ML-based observation operators, in the ECMWF Integrated Forecast System (IFS). The work consists in (1) developing a ML-based observation operator for each type of observation; (2) implement the ML-based observation operators in the IFS to jointly analyse soil moisture and Leaf Area Index (LAI) (3) evaluate the impacts on the forecast of Gros Primary Production and low-level meteorological variables forecast.

The IFS model fields used to predict ASCAT backscatter at 40° include soil moisture and soil temperature in the first 3 soil layers and LAI. A feedforward neural network with 4 hidden layers, 60 neurons is trained over 2016-2018 period and tested over 2019. For SIF, the predictors include LAI, shortwave downwelling radiation, 2m temperature and humidity, soil moisture, root-zone soil moisture, soil temperature and the fraction of low and high vegetation. An XGBOOST model was trained over 2019-2020, tuned over 2021 and tested over 2022. Both SIF and ASCAT ML-observation operators show good performances at global scale. Performances are slightly lower for SIF compared to ASCAT backscatter related to the lack of information content in the IFS model fields to accurately predict the SIF satellite signal at global scale.

The assimilation of SIF provides realistic spatiotemporal patterns of low and high vegetation LAI increments. The updated LAI shows better agreement with the Copernicus satellite LAI product over Northern Eurasia and scattered regions in North and South America, Central Europe and Eastern and Southern Australia. Lower performances are obtained over tropical rainforest and sparse vegetation regions where the prediction of SIF by the ML observation operator is more uncertain. The implementation of the ASCAT observation operator in the IFS is ongoing and results will be presented at the seminar.

# A Multi-Fidelity Ensemble Kalman Filter with a machine learned surrogate model

Jeffrey van der Voort[1], Martin Verlaan[1], Hanne Kekkonen[1]

[1] TU Delft

One of the disadvantages of large physical models is that they can be very computationally expensive. Therefore, ensemble data assimilation approaches, such as the Ensemble Kalman Filter (EnKF), become expensive as they need a large number of ensemble members and thus model runs. In this work we investigate the use of a Multi-Fidelity Ensemble Kalman Filter (MF-EnKF), where the lower fidelity model is a machine learned surrogate model and the high fidelity model is the original full model. The idea behind this is to use an ensemble of a few but expensive full model runs, combined with an ensemble of many cheap but less accurate surrogate model runs. In this way we can reach similar or increased accuracy with less full model runs and thus less computational time. The method is compared to the more traditional multi-fidelity approach in which the lower-fidelity model is a lower resolution model. We investigate the performance by testing the approach on two test problems, namely the Lorenz-2005 model and the Quasi-Geostrophic model. Results show that the MF-EnKF outperforms the EnKF for the same number of full model runs and that the MF-EnKF can reach similar or improved accuracy with less full model runs.