

Finetuning a Weather Foundation Model with Lightweight Decoders for Unseen Physical Processes

Fanny Lehmann^{1,2}, Firat Ozdemir³, Benedikt Soja⁴, Torsten Hoefer⁵,
Siddhartha Mishra¹, Sebastian Schemm⁶

¹Seminar for Applied Mathematics, ETH Zurich, Switzerland

²ETH AI Center, ETH Zurich, Switzerland

³Swiss Data Science Center, ETH Zurich, Switzerland

⁴Institute of Geodesy and Photogrammetry, ETH Zurich, Switzerland

⁵Scalable Parallel Computing Lab, ETH Zurich, Switzerland

⁶Department of Applied Mathematics and Theoretical Physics, University of Cambridge, United Kingdom

Abstract

Recent advances in AI weather forecasting have led to the emergence of so-called “foundation models”, typically defined by expensive pretraining and minimal fine-tuning for downstream tasks. However, in the natural sciences, a desirable foundation model should also encode meaningful statistical relationships between the underlying physical variables.

This study evaluates the performance of the state-of-the-art Aurora foundation model in predicting hydrological variables, which were not considered during pretraining. We introduce a lightweight approach using shallow decoders trained on the latent representations of the pretrained model to predict these new variables. As a baseline, we compare this to fine-tuning the full model, which allows further optimization of the latent space while incorporating new variables into both inputs and outputs.

The decoder-based approach requires 50% less training time and 35% less memory, while achieving strong accuracy across various hydrological variables and preserving desirable properties of the foundation model, such as autoregressive stability. Notably, decoder accuracy depends on the physical correlation between the new variables and those used during pretraining, indicating that Aurora’s latent space captures meaningful physical relationships.

In this sense, we argue that an important quality metric for foundation models in Earth sciences is their ability to be extended to new variables without a full fine-tuning. This provides a new perspective for making foundation models more accessible to communities with limited computational resources, while supporting broader adoption in Earth sciences.

1 Introduction

With the recent developments in machine learning (ML), numerous large deep learning models have been proposed in weather and climate science. Among them, many studies claim to design “foundation models”, without necessarily defining which characteristics of a large pretrained deep neural network defines a foundation model. The rise of foundation models in weather and climate science follows the success of these models in Natural Language Processing (NLP) and Computer Vision (CV) where their definition is less complex than in natural science. For example, Bommasani et al. (2021, p.3) defines a foundation model as “any model that is trained on broad data (generally using self-supervision at scale) that can be adapted (e.g., fine-tuned) to a wide range of downstream tasks”.

Corresponding author: Fanny Lehmann, fanny.lehmann@ai.ethz.ch

The use of broad data is the most prevailing feature in weather and climate foundation models, starting from Climax (Nguyen et al., 2023) that builds on several CMIP6 datasets for pretraining while fine-tuning is performed on ERA5 reanalysis data. This multi-dataset approach was further adopted in ORBIT (Wang et al., 2024) and Aurora (Bodnar et al., 2025), the latter exploring also additional fine-tuning tasks on atmospheric chemistry and ocean wave modelling. Since downstream tasks generally involve fine-tuning with new data and/or a new training objective, it is at this stage not clear if the expensive pretraining phase of a foundation model provides any additional benefit over a task-specific model. The latter would typically consist of less parameters and would be trained specifically on the new objective, e.g., high-resolution prediction (Adamov et al., 2025; Xu et al., 2025), terrestrial water storage prediction (Li et al., 2024; Palazzoli et al., 2025).

In contrast to NLP and CV, self-supervision is seldomly used when pretraining weather foundation models and the focus is typically on forecasting. Nevertheless, some models such as AtmoRep (Lessig et al., 2023) and Prithvi WxC (Schmude et al., 2024) define a self-supervised pretraining task with masking in parallel or in complement to forecasting. Masking is a technique used during the pre-training phase with the aim to reconstruct the randomly masked regions (sometimes masking is also applied across variables). One should also mention WeatherGFM (Zhao et al., 2024) that proposes a “generalist foundation model” to handle several weather understanding tasks by unifying the visual representation of different data sources and tasks. Although applied to satellite data (which prevents comparison with the above-mentioned models), WeatherGFM can realise several downstream tasks with few-shot learning, e.g., super-resolution, weather forecasting, image translation from satellite to radar image.

However, the generic definition of a foundation model pays less attention to the underlying connection between physical variables governed by partial differential equations (PDEs). Although the training data are presented as images, derived from numerical simulations, observations, or a combination of both, they actually represent the values of physical variables, such as temperature and pressure, at specific times. Therefore, we argue that a foundation model in weather science should also capture underlying physical relationships between the predicted variables, thanks to an extensive training on substantial datasets. This point of view can be related to recent developments in designing general purpose foundation models for Physics governed by PDEs (Herde et al., 2024), where it has been shown that pretraining on carefully selected sets of PDEs enables the models to learn rich latent physical representations and provides considerable gains for unseen downstream tasks, over task-specific models.

To provide further insights, this work examines how the latent space, which contains a representation of the multivariate distribution of the training data, can be employed to predict new variables exhibiting possible correlations with those originally included in the training data. This enables us to assess whether the statistical relationship between physical variables has been encoded. Bodnar et al. (2025) demonstrate that the Aurora model can predict new variables by incorporating them as additional inputs and outputs, followed by fine-tuning the entire model for this specific downstream task. In the present work, we instead keep the Aurora model frozen in its pretrained state and learn new variables directly from Aurora’s latent space representation. Following the afore-described conception of a foundation model, the latent space should indeed encode the physical relationships between weather variables, allowing it to transfer physical information to new but related physical processes.

Our approach additionally addresses one limitation of current deep learning weather models, which is their restriction to a limited set of variables (e.g., temperature, pressure, humidity, wind velocity) while many adjacent research and application domains would need more variables for their applications. Hardy and Finney (2025) showed that a Convolutional Neural Network (CNN) can be used to predict 100m wind velocity from the outputs of the AIFS model (which include 10m wind velocity). Since it uses the output

variables, the CNN learns the visual correlations between the different variables and may miss the physical information encoded in the large pretrained latent space. The present work focuses on hydrological variables related to the water cycle (e.g., evaporation, runoff, soil moisture, terrestrial water storage), a field that has not yet been studied with weather foundation models. By learning the new variables directly by decoding the latent space, our approach provides a lightweight extension that alleviates the costs of fine-tuning the entire foundation model.

In the following, Section 2 details the Aurora model, the benchmark models, and the evaluation metrics. Section 3 describes the datasets used, while Section 4 presents the prediction results. Finally, Section 5 discusses the impacts and limitations of our method before drawing some conclusions in Section 6.

2 Methods

All models considered in this work focus on forecasting. They consider inputs at current time t (and previous time $t-\Delta t$, depending on the model) and predict outputs at the next time step $t + \Delta t$, where Δt equals 6 hours if not specified otherwise.

2.1 Aurora Baseline Model

The overall architecture of Aurora (Bodnar et al., 2025) follows an encoder-processor-decoder structure (Fig. 1). The encoder takes as input images of size $T \times H \times W$ (with $T=2$ time steps) that can be static variables, surface variables, and atmospheric variables (in which case one image is given per pressure level). Similar to Vision Transformers (ViT), images are split into patches of size $T \times P \times P$. Surface and static variables are concatenated as image channels V_s . Similarly, atmospheric variables at each pressure level are concatenated as image channels V_a . The patches are then mapped to embedding vectors of dimension E (with $E=512$) via a linear layer $T \times P \times P \times V_* \mapsto E$ (where V_* is either V_s or V_a). The final part of the encoder is a 3D Perceiver (Jaegle et al., 2021) that reduces the pressure levels of the atmospheric variables to a smaller number of 3 latent pressure levels.

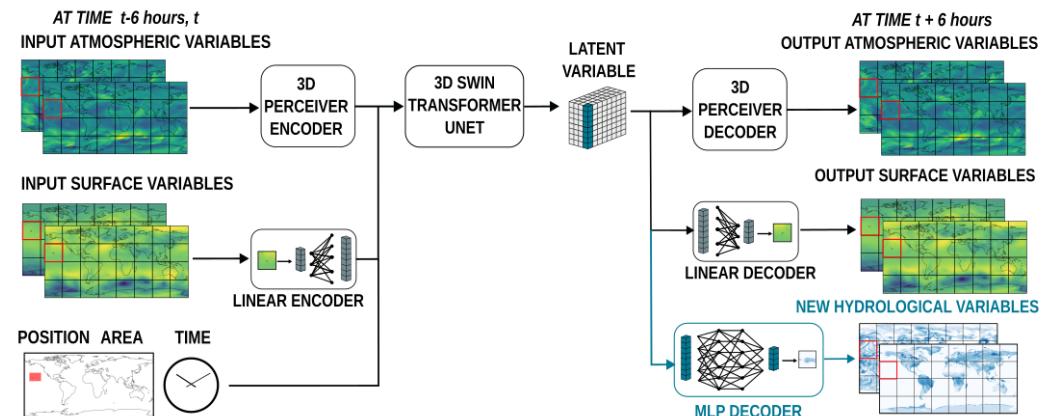


Figure 1: In black, the original Aurora model with its encoder-processor-decoder architecture. These blocks are frozen in our implementation. In blue, the additional Multi-Layer Perceptron (MLP) decoders that are trained to predict new physical variables.

The processor is the core component and is responsible for learning the temporal dynamics. It is based on a 3D Swin Transformer U-Net and the reader is referred to the original work for more details on the architecture. The output of the processor is what we call the latent variable. Its size is $\frac{HW}{P^2} \times 4 \times 2E$, where the first dimension corresponds to the number of patches reshaped along a single dimension, the second dimension is the pressure dimension (3 latent pressure levels + 1 surface level), and the third dimension is the embedding.

Symmetrically to the encoder, surface variables are decoded from first latent pressure level using a linear decoder. The remaining three latent pressure levels are expanded to the number of atmospheric pressure levels with a 3D Perceiver and subsequently decoded into the atmospheric variables with a linear decoder. The decoders of (Bodnar et al., 2025) consist of a single linear layer per variable with $2E$ input neurons and P^2 output neurons, thus transforming the embeddings into a patch.

In our configuration, we use height $H=720$, width $W=1440$, patch size $P=4$, and embedding dimension $E=512$. This corresponds to the pretrained checkpoint available from Bodnar et al. (2025).

2.2 Decoder

To predict new variables, we keep the Aurora model frozen and train one additional decoder per new physical variable. Since we consider only surface variables, we choose a Multi-Layer Perceptron (MLP) with three layers of size E , $E/2$, $E/2$, and ReLU activations in-between. Since the Aurora model is frozen, we marginally increase the representative complexity of the decoders with additional non-linearities compared to the original 1-layer surface decoders. Each MLP decoder contains around 300 000 learnable parameters. In the following, the term *decoders* means that only the MLP decoders are trained.

New hydrological variables include ones that are defined only over land. Accordingly, the loss function for these variables is computed based on a land-sea mask to exclude pixels over the oceans. The loss function is the Mean Absolute Error (MAE) with latitude-dependent weighting. The choice of the MAE follows Aurora implementation and the latitude-dependent weighting is commonly used to account for the geometric distortion of the globe when projected onto a rectangular grid (Chen et al., 2023; Nguyen et al., 2023). Training consists of 10 epochs with a batch size of 8 on 32 GH200 GPUs. The learning rate follows a cosine decay from 5×10^{-4} down to 5×10^{-5} preceded by a linear warmup for 100 steps.

2.3 Benchmark Models

To compare the results of the decoders with other deep learning models, state-of-the-art models that predict at least precipitation are selected. GraphCast is based on a Graph Neural Network (Lam et al., 2023). FuXi builds on Swin Transformer and integrates cascade models for different lead times (Chen et al., 2023). ACE2 relies on Spherical Fourier Neural Operators and adds physical constraints to target long-term stability on climatic time scales (Watt-Meyer et al., 2024). ACE2 predicts more variables than GraphCast and FuXi, in particular variables related to energy fluxes that are used in Section 4.4. ACE2 predictions are available on a $1^\circ \times 1^\circ$ grid and are bilinearly interpolated to $0.25^\circ \times 0.25^\circ$ for comparison with our Decoder (Section 4.4). GraphCast and FuXi predictions are taken from WeatherBench2 (Rasp et al., 2024).

Furthermore, Aurora is fine-tuned for the new physical variables using a standard approach, where new variables serve as input and output. The Aurora baseline architecture is modified only by adding a linear encoder and a linear decoder for each new surface variable. Then, the full 1.3 billion-parameter model is finetuned on a latitude-

weighted MAE loss function that includes all variables. In the loss function, new variables have a weight of 1, while the other variables retain the same weights as in the original Aurora implementation. This model is named Aurora⁺ in the following. Aurora⁺ is fine-tuned on the same dataset as the decoders, for 10 epochs with a batch size of 1 (per GPU), with a learning rate undergoing a linear warmup for 500 steps up to 5×10^{-4} and then a cosine decay.

For the reference numerical predictions, we use the mean of the 50 ensembles from the ECMWF Integrated Forecasting System (IFS), available in WeatherBench2. The IFS model is a previous version underlying the ERA5 reanalysis, which also assimilates observations.

2.4 Evaluation Metrics

The output variable dimension at each time step is $H \times W \times C$ where $C=1$ for surface variables, otherwise C corresponds to the number of atmospheric levels. The agreement between predicted variables \hat{X} and the reference X is evaluated with several metrics. Since this work focuses on surface variables, the metrics definition below omits the third dimension for the sake of simplicity. The MAE

$$\text{MAE}(\hat{X}, X) = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W |\hat{X}(i, j) - X(i, j)| \quad (1)$$

and Root Mean Square Error (RMSE)

$$\text{RMSE}(\hat{X}, X) = \sqrt{\frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W (\hat{X}(i, j) - X(i, j))^2} \quad (2)$$

evaluate the point-wise accuracy. The RMSE gives more penalisation to outliers (such as isolated pixels with significantly lower or higher values than the reference). The bias

$$\text{bias}(\hat{X}, X) = \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W (\hat{X}(i, j) - X(i, j)) \quad (3)$$

is a signed metric that indicates whether predictions are generally underestimated (i.e., bias is negative) or overestimated (i.e., bias is positive). The sample Pearson correlation coefficient (PCC)

$$\text{PCC}(\hat{X}, X) = \frac{\sum_{i=1}^H \sum_{j=1}^W (\hat{X}(i, j) - \bar{\hat{X}})(X(i, j) - \bar{X})}{\sqrt{\sum_{i=1}^H \sum_{j=1}^W (\hat{X}(i, j) - \bar{\hat{X}})^2} \sqrt{\sum_{i=1}^H \sum_{j=1}^W (X(i, j) - \bar{X})^2}} \quad (4)$$

evaluates the linear correlation between the predictions and the reference. A correlation of 1 is perfect, 0 indicates no correlation, and negative values indicate opposite correlation. In Equation 4, $\bar{X} := \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W X(i, j)$ is the average reference, and similarly, $\bar{\hat{X}} := \frac{1}{HW} \sum_{i=1}^H \sum_{j=1}^W \hat{X}(i, j)$ is the average prediction.

The Wasserstein-1 distance (Ramdas et al., 2015) compares the distribution of predicted values to the reference distribution

$$W_1(\hat{X}, X) = \int_{-\infty}^{+\infty} |CDF_{\hat{X}}(x) - CDF_X(x)| dx \quad (5)$$

where CDF_X (resp. $CDF_{\hat{X}}$) is the empirical Cumulative Distribution Function (CDF) computed from samples X_n (resp. \hat{X}_n). For variables with narrow distributions such as precipitation (with a majority of pixels being close to 0), the Wasserstein-1 distance is also computed in the log space, $W_1(\log(\hat{X}), \log(X))$.

For precipitation specifically, two categorical metrics are introduced to capture the ability of the decoder to predict values above a given threshold. The Fraction Skill Score (FSS, Roberts and Lean (2008)) evaluates the fraction of pixels above a threshold α (with $\alpha=1$ mm and $\alpha=5$ mm in this work) in windows of size 11 pixels \times 11 pixels. The FSS is defined as

$$\text{FSS}(\hat{X}, X) = 1 - \frac{\sum_w (f_w(\hat{X}) - f_w(X))^2}{\sum_w f_w(\hat{X})^2 + \sum_w f_w(X)^2} \quad (6)$$

where $f_w(X)$ (resp. $f_w(\hat{X})$) is the fraction of pixels above threshold α in the reference (resp. predicted) field and w iterates over all windows in the $H \times W$ field. The Stable Eq-uitable Error in Probability Space (SEEPS, Rodwell et al. (2010)) classifies precipitation into three categories: ‘dry’ when 6-hour precipitation is lower than 0.25 mm, ‘light rain’ between 0.25 mm and a *wet threshold* determined from climatology, and ‘heavy rain’ above the *wet threshold*.

3 Data

The main training dataset is the ERA5 reanalysis, widely used to train AI weather models thanks to its high quality and long time span (Hersbach et al., 2020). Aurora uses five atmospheric variables on 13 pressure levels: temperature, specific humidity, geopotential, northward and eastward wind speed; four surface variables: 2-metre temperature, mean sea level pressure, northward and eastward 10-metre wind velocity; and three static variables: land-sea mask, geopotential at the surface, and the soil type. All these variables come from ERA5 and are given as inputs to Aurora to get the latent vector corresponding to each time step.

In addition, the training targets for the decoders include 6-hour accumulated potential evaporation, runoff (the sum of surface and subsurface runoff), and soil moisture (the sum of volumetric soil water layer weighted by the layer thickness for the first three layers defined in ERA5, i.e., down to 1-metre depth). While our main focus is on hydrological variables, we also add variables related to the surface and top of atmosphere energy balance: surface latent heat flux, surface sensible heat flux, surface net solar radiation, surface net thermal radiation, surface solar radiation downwards, surface thermal radiation downwards, top net solar radiation, and top net thermal radiation. All energy variables are 6-hour accumulations at T00, T06, T12, and T18 UTC, computed from the 1-hourly data given in ERA5 (the sum from T01 to T06 is chosen for the accumulation at T06, and similarly for T12, T18, T00).

Due to known biases for precipitation in ERA5, the Multi-Source Weighted-Ensemble Precipitation version 2.80 (MSWEP) dataset is used instead (Beck et al., 2019). MSWEP is a combination of gauge-, satellite-, and reanalysis-based data providing global coverage of total precipitation at a 0.1° spatial resolution, 3-hourly temporal resolution, from 1979 to 2022. MSWEP was found to be more accurate than ERA5 for hydrological modelling (Gebrechorkos et al., 2024). This work uses a bilinear spatial interpolation to regrid MSWEP to 0.25° and sums two consecutive samples to obtain 6-hour accumulated precipitation. Precipitation is log-transformed as $\log(1 + x/\epsilon)$ where $\epsilon = 10^{-5}$, following Rasp et al. (2020).

The training dataset comprises 17 000 samples between May 2002 and June 2014, while the validation dataset comprises 1700 samples between July 2014 and December 2015, and the test dataset contains 1460 samples between January 2020 and December 2020.

Table 1: Metrics of our decoder (first column) and our Aurora⁺ predictions (second column) for the new hydrological variables. Predictions are done with a 6-hour lead time in 2020. Precipitation metrics are given in Tab. 3. The metric definitions are given in Section 2.4. Better results are indicated in **bold**.

Variable	Metric	Decoder	Aurora ⁺
potential evaporation	MAE (mm)	0.040	0.019
	RMSE (mm)	0.177	0.079
	bias (mm)	-0.003	0.003
	PCC	0.958	0.992
	W1	0.007	0.005
runoff	MAE (mm)	0.005	0.003
	RMSE (mm)	0.087	0.079
	bias (mm)	-0.003	-0.002
	PCC	0.420	0.559
	W1	0.004	0.002
soil moisture	MAE (m)	0.015	0.001
	RMSE (m)	0.034	0.003
	bias (m)	-0.002	0.001
	PCC	0.969	0.999
	W1	3.181	0.578

4 Results

4.1 Prediction of Hydrological Variables

Figure 2 illustrates the predictions of four hydrological variables with the decoders on July 4th, 2020: total precipitation, potential evaporation, runoff, and soil moisture. First, a visual inspection of Fig. 2 indicates that the decoder prediction is close to the baseline in many key locations. This is especially remarkable given that hydrological variables exhibit visual patterns very different from the original atmospheric variables (Fig. B1 vs. Fig. 2). As expected from MAE-based training, regions of high amplitude or strong gradients present the greatest challenges for the decoders. Overall, the decoders, despite their simple architectures, capture the visual patterns of the new variables.

Accuracy metrics are quantified for all 1460 test samples in 2020 in Tab. 1 and Tab. 3. Metrics confirm that, overall, decoders are able to predict all hydrological variables under study. Good accuracy is observed for potential evaporation and soil moisture, with correlation coefficients above 0.95 and relative MAE of 0.25 (resp. 0.17) for potential evaporation (resp. soil moisture). The energy spectra confirm the good agreement at all spatial scales, with a slight underestimation by the decoders for small spatial scales (Fig. C1). Runoff, however, shows higher errors (relative MAE of 0.66) and low correlation coefficients (0.42). Precipitation is discussed in detail in the following Section 4.2.

The accuracy obtained with Aurora⁺ is listed in Tab. 1 to give an indication of the error that can be obtained when fine-tuning the full foundation model for new variables. However, we want to emphasise again that the goal is not to determine the best model between the decoders and Aurora⁺, given the significant difference in required training effort (Tab. 2). The accuracy of the lightweight decoder closely matches the optimal Aurora⁺ for potential evaporation, while a larger difference is observed for soil moisture. This indicates that the latent space captured most of the physical information necessary for predicting potential evaporation, whereas soil moisture benefits from additional learning.

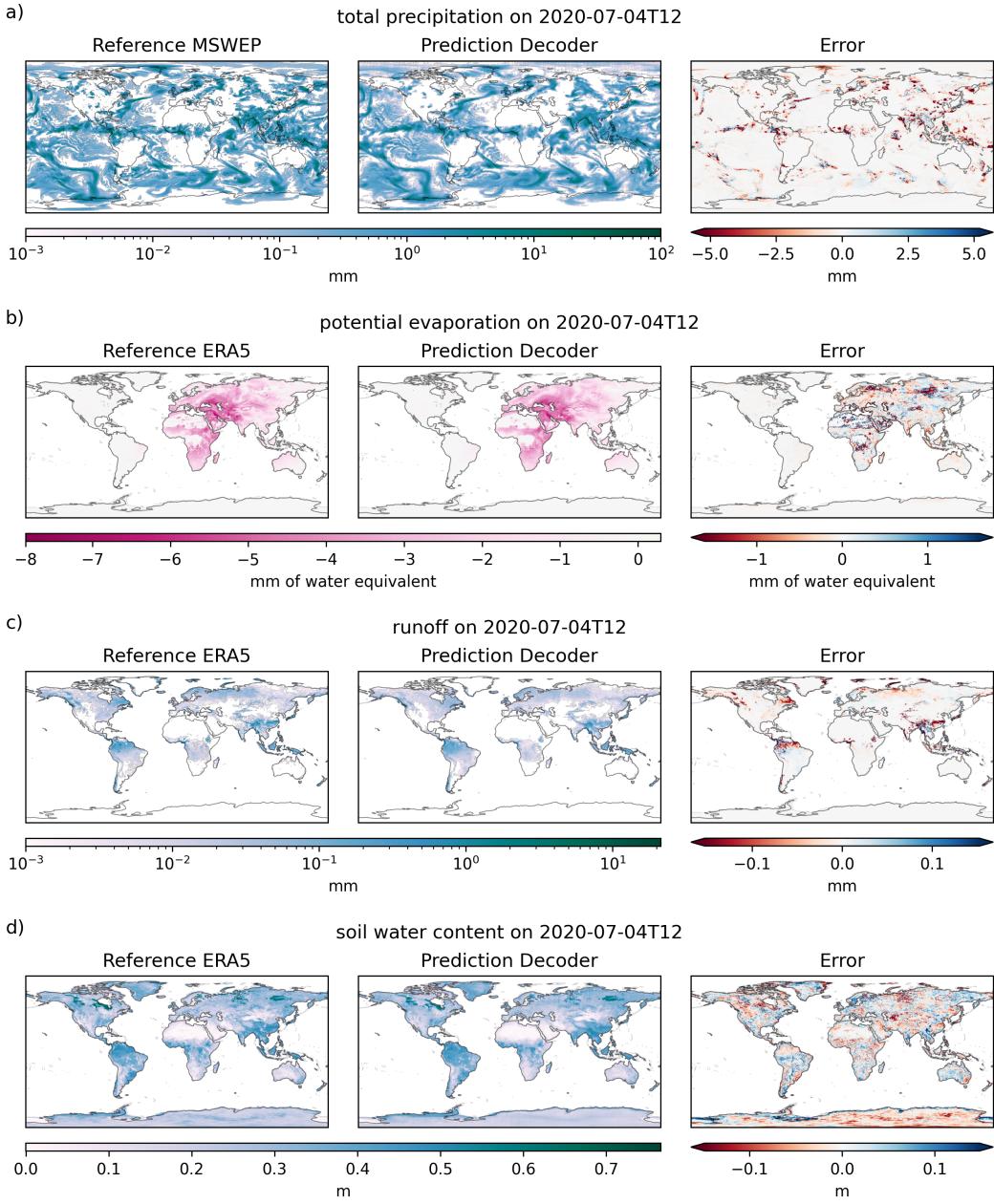


Figure 2: Next-step prediction at $t+6h$ of a) total precipitation [mm/6hr], b) potential evaporation [mm/6hr], c) runoff [mm], d) soil moisture [m]. Left column is the reference, middle column is the prediction from the decoder, right column is the point-wise error between the prediction and the reference on July 4th, 2020.

This can be understood from the fact that potential evaporation is mostly determined by wind velocity, air temperature, and pressure, all variables included in the pre-training of Aurora. In contrast, soil moisture also depends on land surface processes absent from the original Aurora model, making it difficult to decode from the latent space using MLPs.

Table 2: Training speed evaluated by the number of Floating Point Operations Per Second (FLOPS) and the number of samples per second. Training memory requirements quantified with the allocated GPU memory.

	Decoder	Aurora ⁺
FLOPS	4×10^{11}	31×10^{12}
samples/sec	0.34	0.16
allocated GPU memory	65GB	99GB

Significant gains in training speed and GPU memory requirements compensate for the decreased accuracy of the decoders compared to Aurora⁺. The decoders can process 0.34 samples per second, compared to 0.16 for Aurora⁺, making their training twice faster (Tab. 2). The peak GPU memory necessary to train the decoders is also significantly reduced, from 99 GB to 65 GB.

4.2 Precipitation Prediction and Comparison with Benchmarks

Precipitation is generally considered as a challenging variable to predict due to its high spatial and temporal variability. Nevertheless, precipitation maps in Fig. 2a illustrate a good agreement between the prediction and the reference data. The regions affected by the largest errors can be roughly divided into two groups. Along the fronts of extratropical cyclones in extratropical latitudes, the predicted values are too low. In the tropics, the largest errors occur in regions of active convection along the equatorial belt.

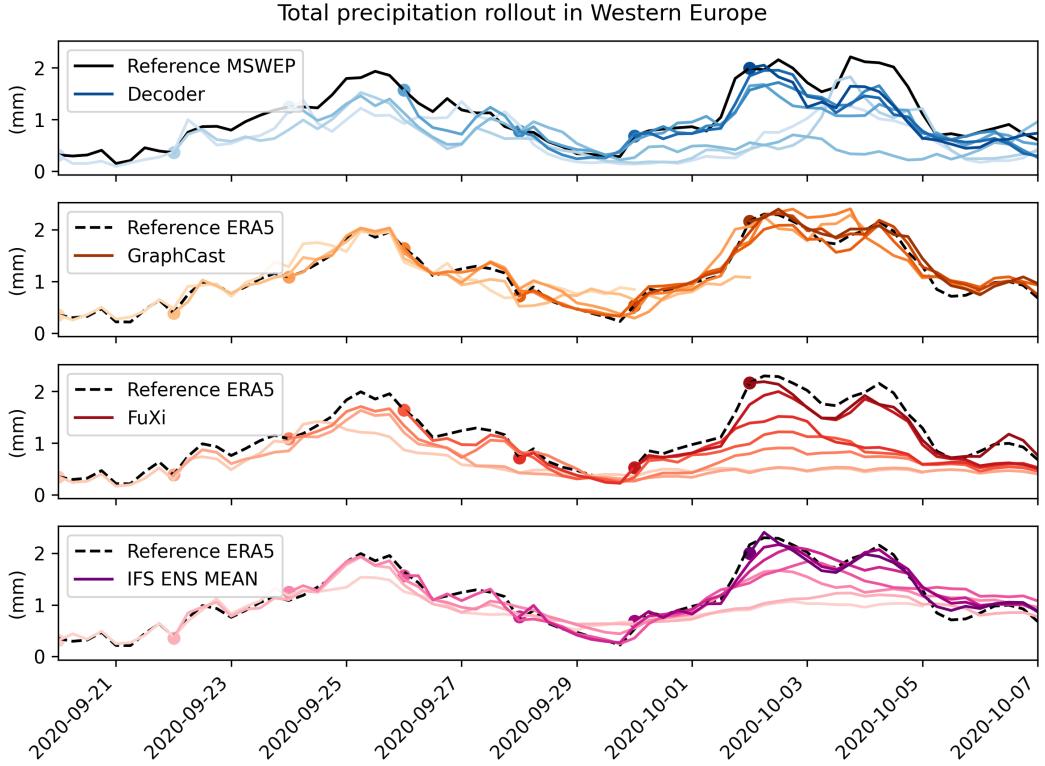
The metrics in Tab. 3 show that the decoder achieves a good accuracy when predicting precipitation. FSS evaluates the ability to predict large precipitation values: 0.92 for a threshold of 1 mm per 6 hour and 0.82 for 5 mm per 6 hour. The correlation between the decoder prediction and the reference MSWEP dataset is satisfying (0.71).

The fact that the decoder is skilful at predicting precipitation should be analysed in the light of the physical correlation between precipitation and atmospheric variables learnt by the original Aurora. In particular, it is known that regions of strong moisture flux convergence often coincide with precipitation and the convergence of moisture flux is the product of wind velocity and specific humidity. Therefore, despite the variability of precipitation and the visual difference with the original Aurora variables, the encoded statistical relationship between the atmospheric variables makes precipitation a learnable variable.

As a reference, accuracy metrics on two benchmark AI models (GraphCast and FuXi) are also provided in Tab. 3, alongside IFS numerical predictions. Results generally indicate better metrics for our decoder and Aurora⁺. However, it should be noted that the reference dataset is MSWEP, which is not the training objective of GraphCast and FuXi. Setting ERA5 as the precipitation reference (while still training on MSWEP) degrades the metrics of the decoder and Aurora⁺ but keeps them comparable to IFS (Tab. D1).

Next, consideration is given to the extreme precipitation associated with the extratropical cyclone “Alex” affecting Europe in early October 2020 (mainly Southern France and Northern Italy). Time series in Fig. 3 indicate that the decoder is able to predict

(a)



(b)

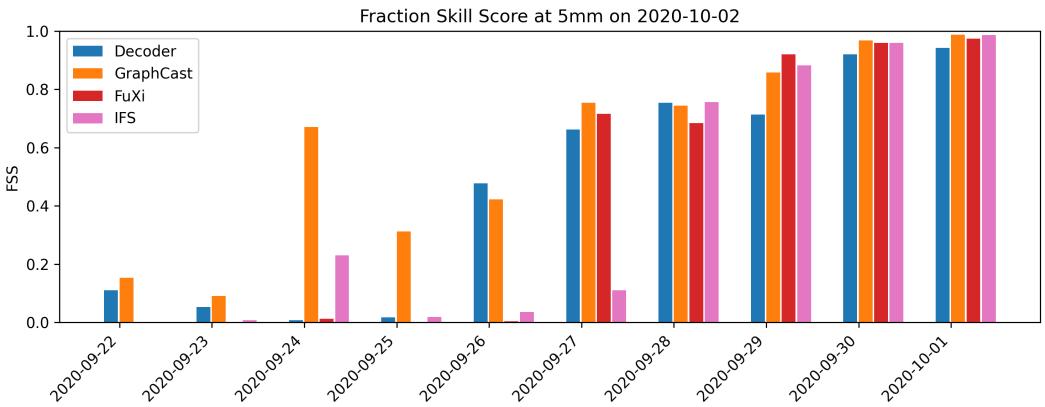


Figure 3: (a) Average of precipitation prediction in Western Europe (area shown in Fig. 4) for rollouts initialised every two days between 2020-09-20 and 2020-10-02 at T00. Each rollout is indicated with a different color shade (earlier rollouts with more pale colors) and the initial time is depicted with a dot. Each panel shows predictions with different models: our decoder (1st row), GraphCast (2nd row), FuXi (3rd row), IFS ensemble mean (4th row). (b) FSS with a 5mm threshold between observed and predicted precipitation on 2020-10-02T06. Predictions are obtained with rollouts initialized between 2020-09-22T00 and 2020-10-01T00. GraphCast, FuXi, and IFS are evaluated against ERA5 while our decoder is evaluated against MSWEP.

Table 3: Precipitation metrics of our decoder (first column) and our Aurora⁺ predictions (second column). Predictions are done with a 6-hour lead time in 2020. The reference dataset is MSWEP. See Tab. D1 for the same metrics with ERA5 as the reference dataset. The best metrics are indicated in **bold**, the second best are underlined.

	Decoder	Aurora ⁺	IFS	GraphCast	FuXi
MAE (mm)	0.32	0.22	0.44	0.38	0.36
RMSE (mm)	<u>1.40</u>	1.03	1.60	1.50	1.42
SEEPS	0.35	0.23	0.34	0.33	<u>0.31</u>
W1	0.246	0.17	0.06	<u>0.15</u>	0.18
W1 log	0.046	<u>0.034</u>	0.029	0.047	0.042
bias (mm)	-0.14	<u>-0.096</u>	0.10	-0.02	-0.02
FSS 1mm	<u>0.92</u>	0.97	0.89	0.90	0.91
FSS 5mm	<u>0.82</u>	0.93	<u>0.82</u>	0.80	0.81
PCC	<u>0.71</u>	0.86	0.68	0.70	0.71

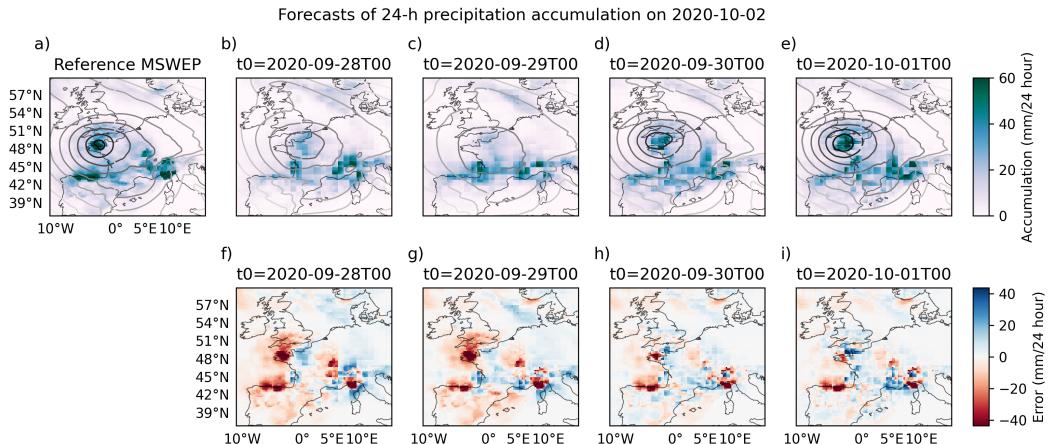


Figure 4: 24-hour accumulation of precipitation in Western Europe on October 2nd, 2020, predicted by our decoder with a rollout from b) 2020-09-28 (18 time steps), c) 2020-09-29 (14 time steps), d) 2020-09-30 (10 time steps), e) 2020-10-01 (6 time steps). Panel a) shows the reference 24-hour accumulation on October 2nd, 2020. The bottom row (f-i) shows the error between the prediction and the reference for each rollout.

the large precipitation values averaged over the region (11°W , 15°E , 37°N , 60°N) up to six days in advance. The longest lead time (18 6-hourly time steps) allows only anticipating the location of the three precipitation hot spots (Fig. 4b). The two regions in northern Spain and France, in particular, are too weak. Decreasing lead time improves the forecast location and intensity of intense precipitation, although the intensity is often too weak. The 6-hour accumulated precipitation observed on October 2nd at T00 was predicted by our decoder with a 5 mm FSS of 0.94 (resp. 0.92) one day (resp. two days) in advance. The underestimation of extreme events amplitude is in agreement with observations from various ML weather forecasting models (Liu et al., 2024).

4.3 Evaluation of Temporal Dynamics

A spatial scale of interest in hydrology is the river basin (or catchment area), which is the area where water flows towards the river of the same name. Many hydrological stud-

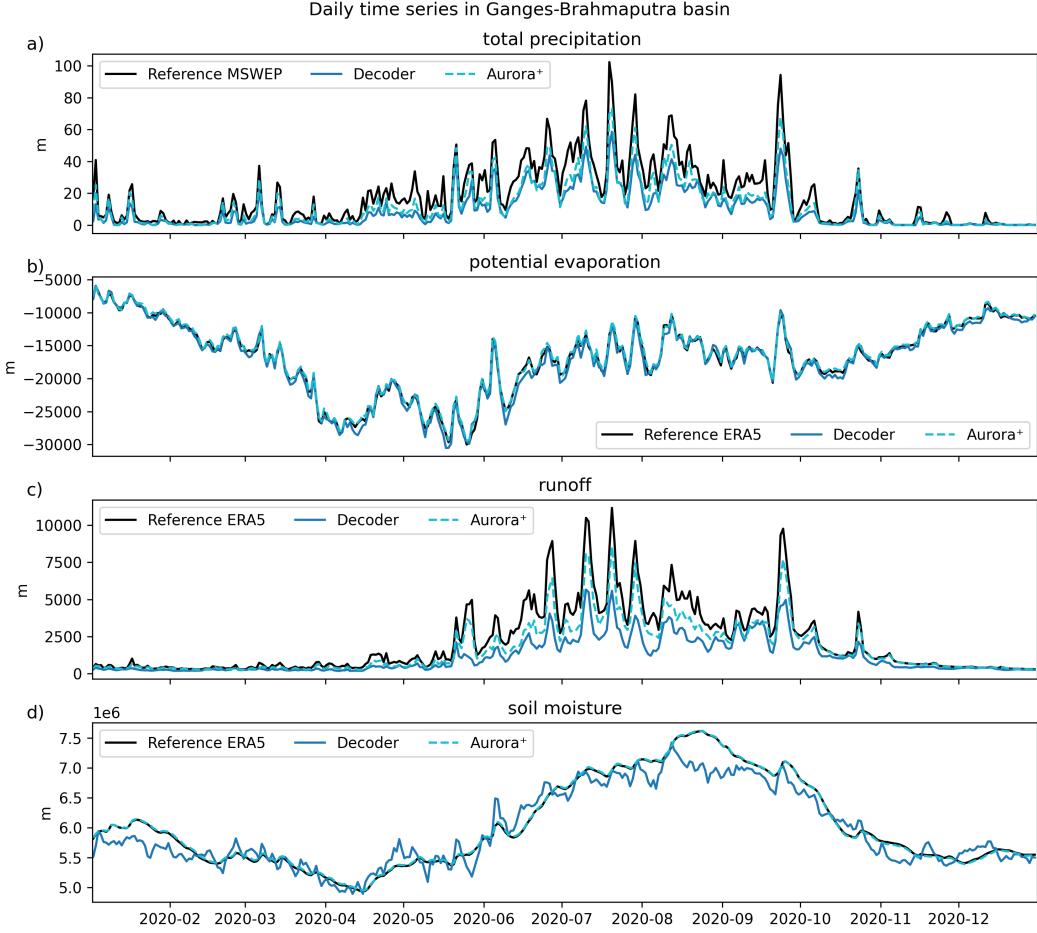


Figure 5: Daily average of 6-hour time series of a) total precipitation [m/6hr], b) potential evaporation [m of equivalent water/6hr], c) runoff [m], d) soil moisture [m] in the Ganges-Brahmaputra basin (shown in red in Fig. 6). Time series are obtained as the weighted sum of pixels inside the basin.

ies aim at estimating hydrological variables from weather predictors, similar to the outcomes provided by our decoders. We computed hydrological time series in 24 major river basins¹ (Fig. 6) by summing the contribution of each pixel inside the basins (with weighting proportional to the latitude-dependent pixel area). Results in the Ganges-Brahmaputra basin are illustrated in Fig. 5, with daily averages for an easier visualisation.

One can observe that the decoders reproduce the temporal dynamics of all hydrological variables for all seasons (Fig. 5). As already noted from the metrics in Tab. 2, potential evaporation shows excellent agreement with the reference, and the main temporal fluctuations of soil moisture are also correctly predicted by the decoders. Due to the slow soil moisture dynamics, an increment of 6 hours leads to very small changes. Nevertheless, despite training only with a 6-hour lead time, our decoder still captures

¹ Amazon, Congo, Mississippi, Ob, Nile, Parana, Yenisey, Lena, West Sahara, Amur, Niger, Ganges-Brahmaputra, Yangtze, Mackenzie, Volga, Zambezi, Lake Eyre, Murray, Nelson, Saint Lawrence, Orinoco, Yellow river, Orange, Danube. Basin shapes were extracted from the level 3 of the HydroBasins dataset (Lehner & Grill, 2013).

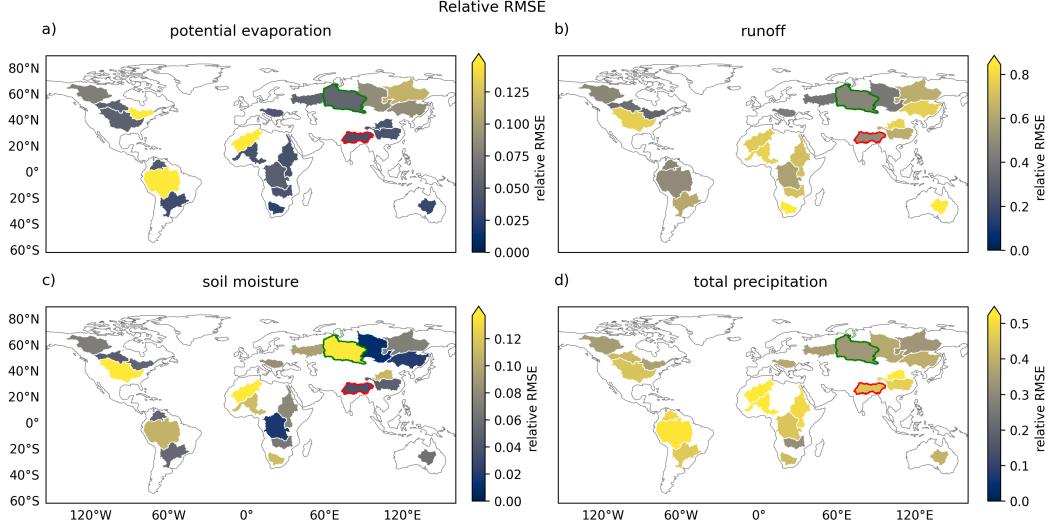


Figure 6: Relative RMSE computed on daily sums at the scale of a river basin. Highlighted in red is the Ganges-Brahmaputra basin, whose detailed time series are given in Fig. 5, in green the Ob basin, with corresponding time series in Fig. E1.

the seasonal change. However, since the temporal dynamics of the soil moisture decoder is driven by the latent space, the decoder leads to soil moisture variations that are faster than the reference. In contrast, fine-tuning the entire foundation model correctly reproduces the slow temporal dynamics of soil moisture (Fig. 5d).

The overall evolution of runoff shows good agreement with the reference, but the decoder (and to a lesser extent, Aurora⁺) underestimates the total runoff (Fig. 5). The runoff peaks between June and October, correlated with high precipitation events, are especially noticeable. The decoder correctly identifies these peaks, but they are notably too weak compared to the reference and the fine-tuned Aurora⁺. This reinforces the importance of the relationship between atmospheric and land surface processes for predicting the latter from the latent space. Runoff in the Ob basin, located in a cold climate, is largely driven by snow melt between April and September, a process the decoder cannot predict (Fig. E1). However, Aurora⁺ learns it, indicating that fine-tuning is superior in this case.

4.4 Prediction of Energy Fluxes

Although the main focus of this work is on surface hydrological variables, decoders with the same architecture are also trained to predict a different set of variables, namely variables linked with energy fluxes. This allows studying the connection between the latent space and energy fluxes measured on the surface and at the top of the atmosphere. The decoders predict eight variables related to energy fluxes (in a training independent from the hydrological variables): surface latent/sensible heat flux, surface net solar/thermal radiation, surface solar/thermal radiation downwards, top net solar/thermal radiation (noting that the two latter top-of-atmosphere variables are not predicted by ACE2).

Rollouts were initialised on the first day of each month from January 2020 to December 2020 and conducted for 384 hours (i.e., 16 days, 64 steps). Figures 7 and 8 show that the decoders are able to predict all variables related to energy fluxes. The decoders yield RMSE close to those obtained with ACE2 for heat fluxes and thermal radiations, both for short and long lead times. Surface solar radiation presents higher errors with

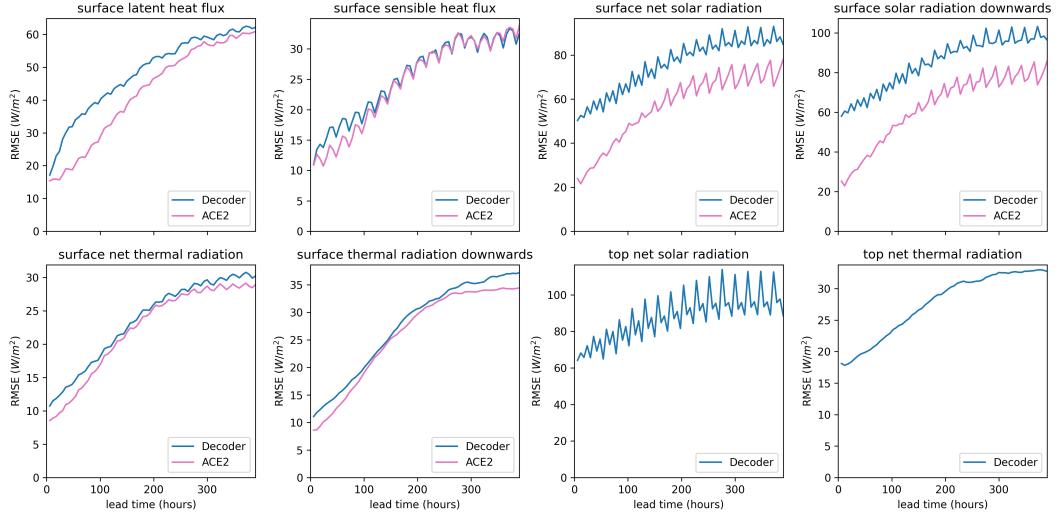


Figure 7: Root Mean Square Error (RMSE) of eight energy variables computed for lead times 6 hours to 384 hours (i.e., 16 days) with our decoder (blue) and ACE2 (pink). Errors should not be compared between variables as they take different values.

the decoder than ACE2, which can be understood by examining the maps in Fig. 8. Indeed, solar fluxes concentrate on the region of the globe that receives the most solar radiation at that time of day, leaving half of the map without radiation (Fig. 8a). These patterns differ from the original Aurora variables, whereas thermal radiation (Fig. 8b) more closely resembles the type of fields already encoded in Aurora, making it easier to predict. ACE2 being fully trained with energy variables as target objectives yields accurate predictions for all energy variables. Figure 7 also shows that the decoder can learn variables at the top of the atmosphere (top net solar/thermal radiation), not only surface variables.

5 Discussion and Limitations

New variables representing physical processes at the surface or at the top of the atmosphere can be learnt by the decoders. Despite their simple architecture, the decoders demonstrate skill across a wide range of variables, from hydrology to energy fluxes, and for different time scales. This shows that the latent space of a foundation model provides the appropriate input encoding to extend the model skills to new downstream tasks. This corroborates previous findings on the Poseidon PDE foundation model that was trained only on equations from fluid dynamics and could be fine-tuned on different physics (i.e., the reaction-diffusion Allen-Cahn equation) with a frozen latent space (Herde et al., 2024).

The decoders also inherit the properties of the underlying foundation model, such as the auto-regressive stability. This allows the decoders to be readily applied for roll-outs without the need to fine-tune them specifically for this objective. Such capability proved advantageous in anticipating the intense precipitation associated with the extra-tropical cyclone “Alex”. Although the decoders depend solely on the rollout accuracy of Aurora, the rate at which prediction errors increase with longer lead times is similar to that of ACE2, a task-specific model fully trained with a rollout objective.

Thanks to their lightweight architecture, training the decoders is twice as fast as fine-tuning the entire foundation model. During the backward pass, gradients are computed only for the decoders, leading to small computational graphs. This allows train-

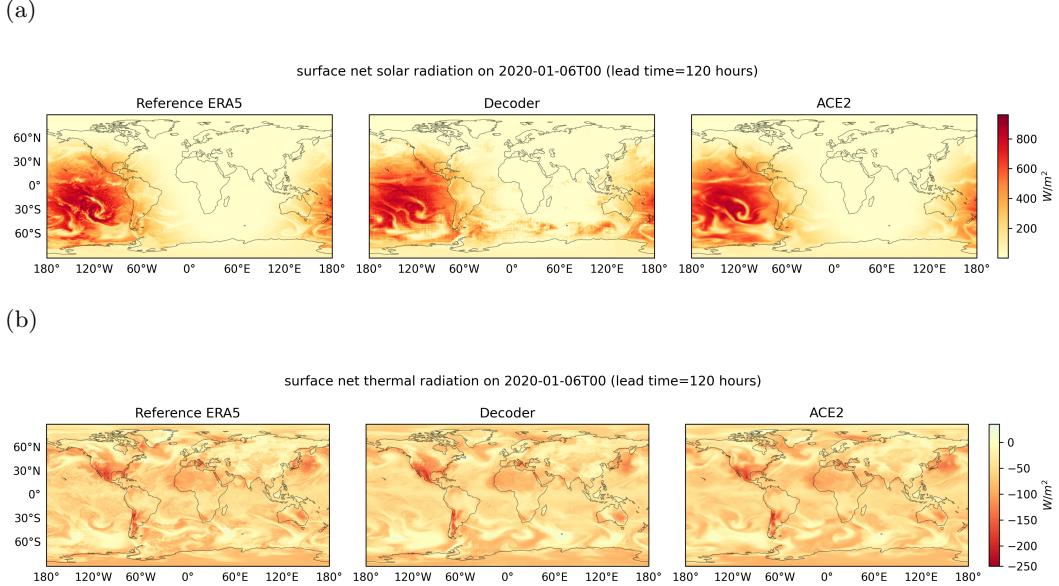


Figure 8: Map of a) surface net solar radiation and b) surface net thermal radiation (bottom panels) from the reference ERA5 (left), predicted by our decoder (middle), and predicted by ACE2 (right).

ing without advanced checkpointing strategies, making them more accessible to a wide scientific community. The decoders can thus be considered as a *frugal* extension of the foundation model.

The accuracy of the decoders is influenced by the physical correlation between atmospheric processes learnt by the original Aurora, whose dynamics have therefore been encoded in the latent space. This was observed for precipitation, for instance, where the decoder achieves accurate predictions. In addition, the accuracy relates to the visual similarities between the original and new fields, as illustrated by the remaining artefacts on solar radiation (Fig. 8a). The decoders cannot learn variables having little connection with atmospheric processes. This was observed, for instance, with terrestrial water storage, which encompasses all water bodies, from the surface to groundwater (Fig. B2).

The decoders exhibit some limitations with high-resolution prediction, as can be seen in Fig. D2. The precipitation map shows some ‘‘patchy’’ artefacts with squares of 4 pixels×4 pixels. These correspond to the patches created for the Swin transformer processor in the pretrained Aurora model. For variables with complex spatial variability, the decoders struggle to learn the correct values for each pixel inside the patch and resort to predictions at the patch scale. However, it is important to note that the predicted patches are not constant and the ‘‘patchy’’ artefacts are much less visible for smoother variables (e.g., soil moisture), as illustrated in Fig. F1.

6 Conclusion

This study investigates the predictions of new surface variables from the latent representations of a weather foundation model using only lightweight decoders and without fine-tuning the entire model. It can be concluded that new hydrological and energy variables can be predicted fairly accurately using these shallow MLP decoders. The de-

coders capture the spatial patterns of the new variables and their variations on different time scales, from sub-daily fluctuations to seasonal change.

The accuracy of the decoders improves when the target physical processes exhibit a close correlation with variables that have been encoded during pretraining. This indicates the effectiveness of the latent space of the Aurora foundation model in encoding statistical relationships between physical processes.

Although ERA5 is the main data source in this work, precipitation is taken from the MSWEP dataset to avoid known biases in ERA5 precipitation. The precipitation accuracy is satisfying with the decoders and the fine-tuned Aurora⁺ model, illustrating that different data sources can be used during training. This allows a selection of the most appropriate datasets from *a priori* knowledge.

The decoders are competitive with benchmark models, although their accuracy does not exceed that of fully fine-tuned AI weather models. This is expected since only a small part of the overall model is trained. Furthermore, forecasts at time $t + \Delta t$ do not rely on target variable observations at time t , which are not available at input. On the other hand, their main benefit is the considerably reduced computational and memory costs. These advantages are potentially significant for startup companies and small research labs without access to substantial High-Performance Computing (HPC) systems.

Our findings indicate that rich latent space representations allow accurate predictions of new variables with lightweight extensions. This calls for developing future foundation models that encompass a large number of physical processes, such that their latent space readily encodes correlations between diverse physical processes. Such foundation models would allow a broad adoption by diverse sectors with various needs and encourage a frugal use of AI resources.

Appendix A Training Details

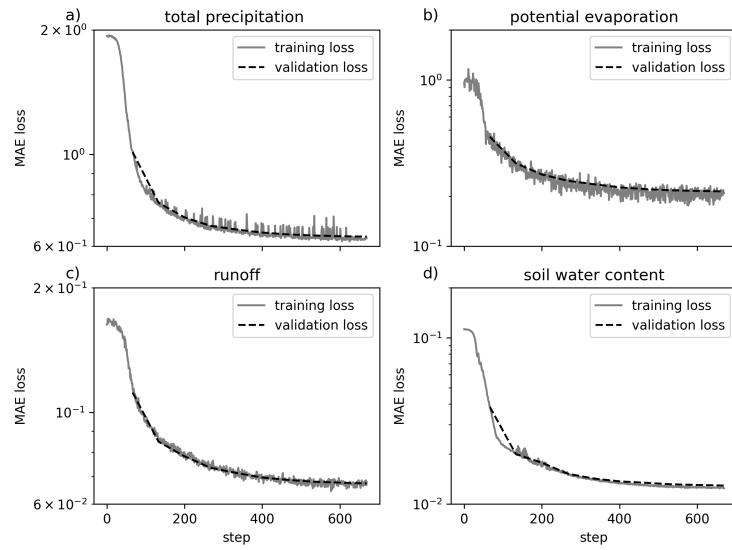


Figure A1: Training loss (grey line) and validation loss (dashed black line) for the decoders

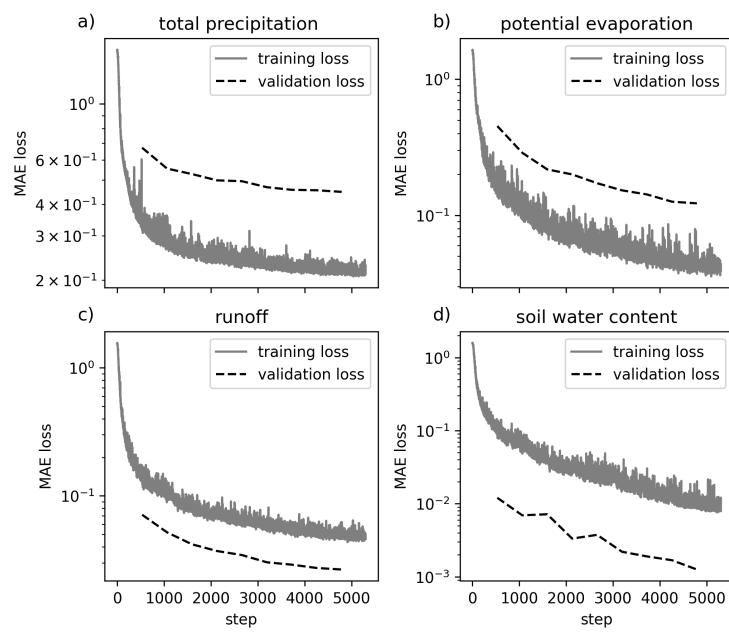


Figure A2: Training loss (grey line) and validation loss (dashed black line) for Aurora⁺

Appendix B Prediction of Surface Variables

In Fig. B2, terrestrial water storage is taken from (Gou & Soja, 2024) and linearly interpolated to 6-hour time steps.

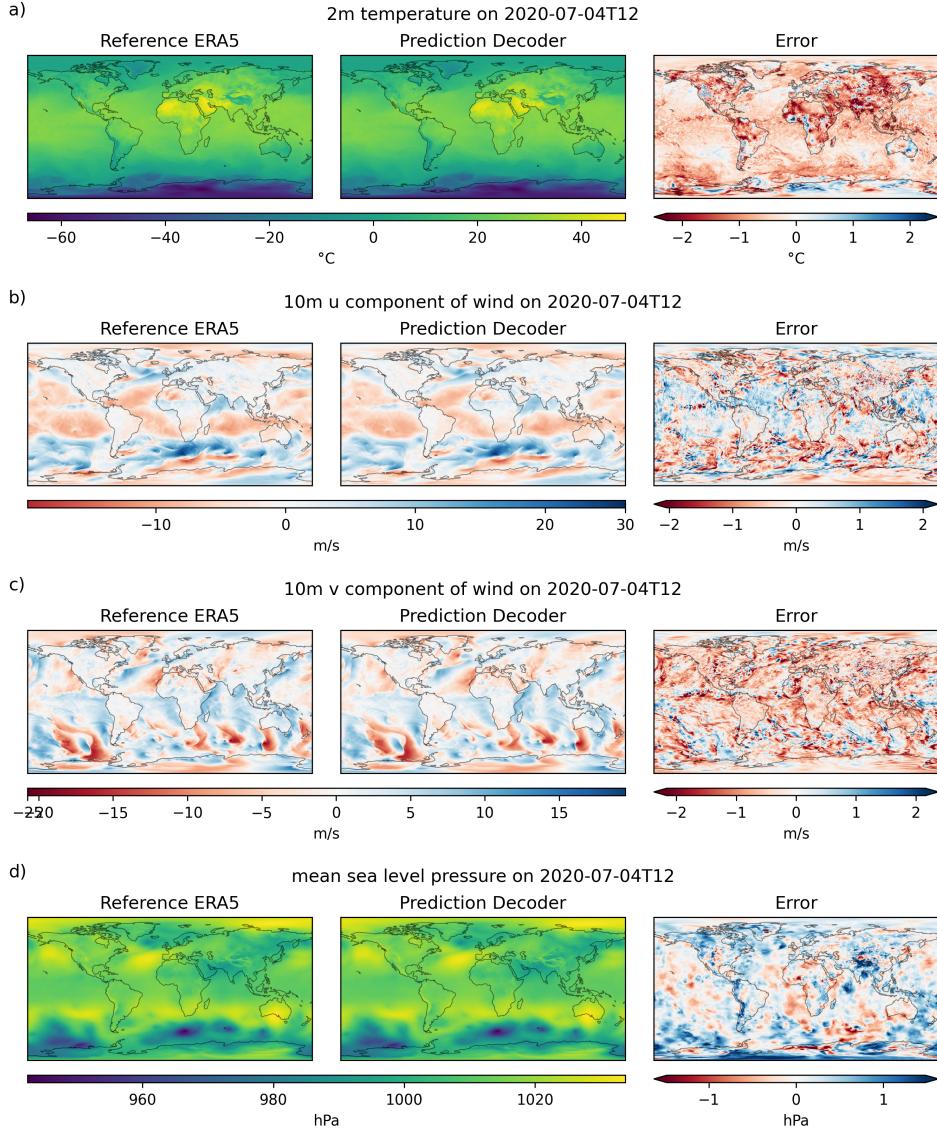


Figure B1: Next step prediction of a) 2-metre temperature, b) 10-metre eastward wind velocity, c) 10-metre northward wind velocity, d) mean sea level pressure. Left column is the reference, middle column is the prediction from frozen Aurora, Right column is the point-wise error

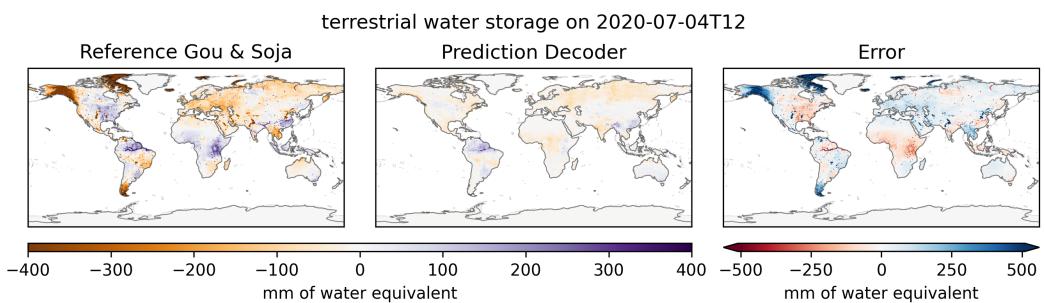


Figure B2: Next-step prediction at t+6h of terrestrial water storage. Left column is the reference, middle column is the prediction from the decoder, right column is the point-wise error between the prediction and the reference on July 4th, 2020.

Appendix C Energy spectra

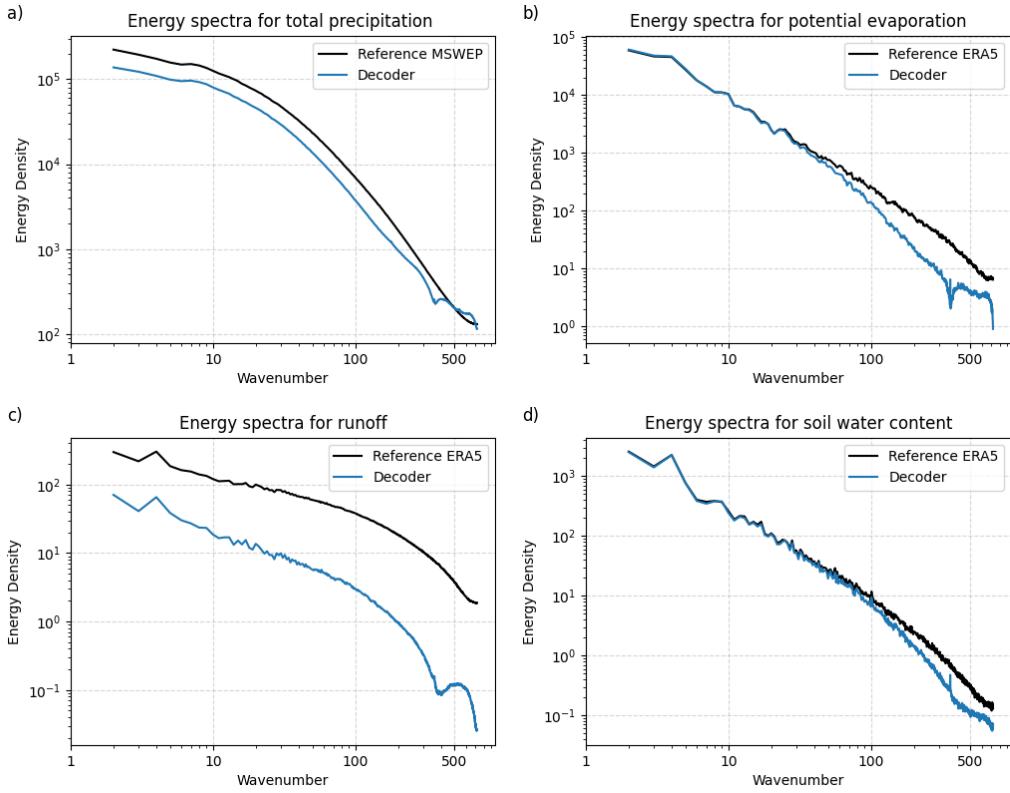


Figure C1: Energy spectra of a) total precipitation, b) potential evaporation, c) runoff, d) soil moisture. Spectra are computed along the longitude dimension for latitude-weighted variables and averaged over the time dimension.

Appendix D Additional Results Precipitation

Table D1: Precipitation metrics of our decoder (first column) and our Aurora⁺ predictions (second column). Predictions are done with a 6-hour lead time in 2020. The reference dataset is ERA5 instead of MSWEP used in Tab. 3

	Decoder	Aurora ⁺	IFS	GraphCast	FuXi
MAE (mm)	0.388	0.306	0.379	0.220	0.191
RMSE (mm)	1.693	1.341	1.629	0.828	0.831
SEEPS	0.400	0.297	0.288	0.234	0.199
W1	0.276	0.207	0.068	0.183	0.222
W1 log	0.038	0.026	0.012	0.053	0.043
bias (mm)	-0.162	-0.124	0.076	-0.046	-0.049
FSS 1mm	0.885	0.924	0.928	0.970	0.982
FSS 5mm	0.725	0.846	0.854	0.941	0.954
PCC	0.622	0.800	0.695	0.922	0.928

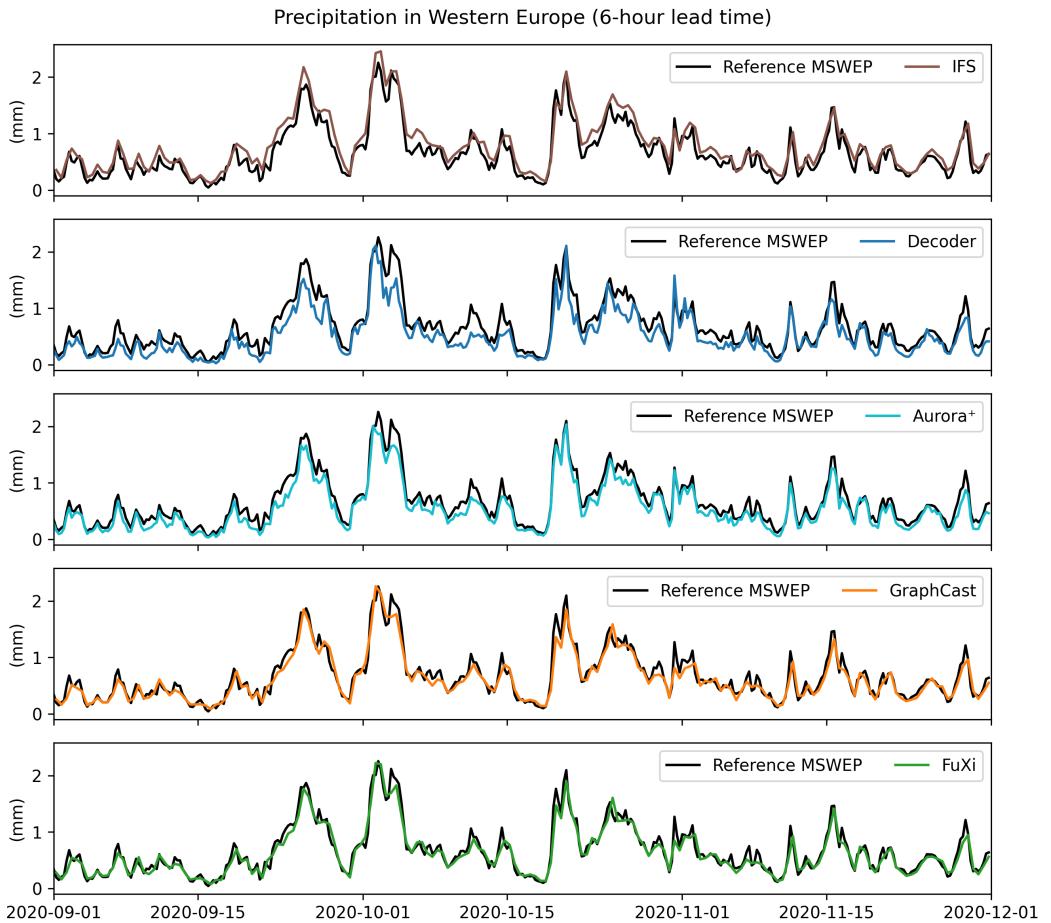


Figure D1: Precipitation prediction during the extratropical cyclone Alex (early October) with a 6-hour lead time from IFS ensemble, our decoder, our Aurora⁺, GraphCast, FuXi. Time series show the spatial average in a box encompassing Western Europe (i.e., the mean over all points in the red box shown in Fig. D2).

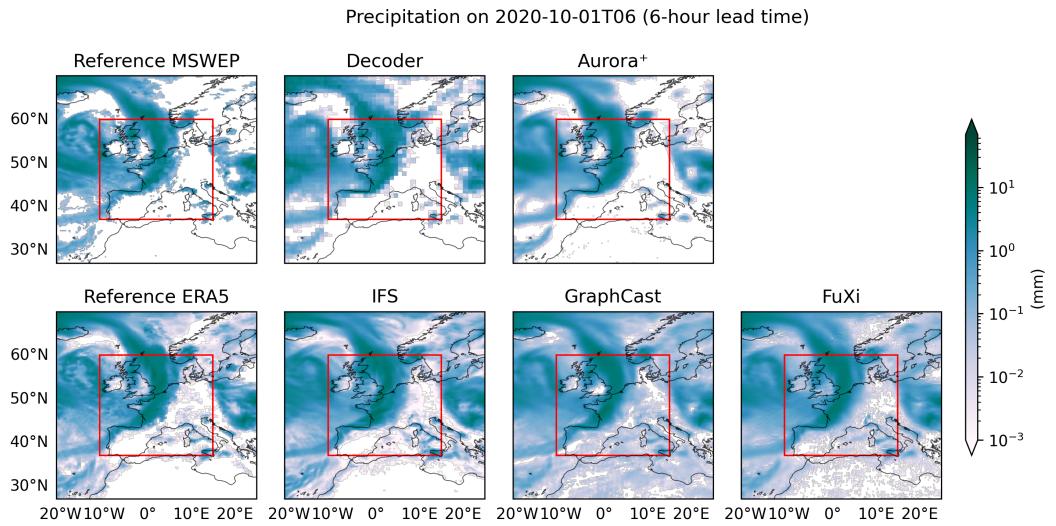


Figure D2: Precipitation in Western Europe on October 1st, 2020, predicted with a 6-hour lead time from IFS ensemble, our decoder, our Aurora⁺, GraphCast, FuXi. The red box shows the averaging region in Fig. D1.

Appendix E Additional Results Temporal Dynamics in River Basins

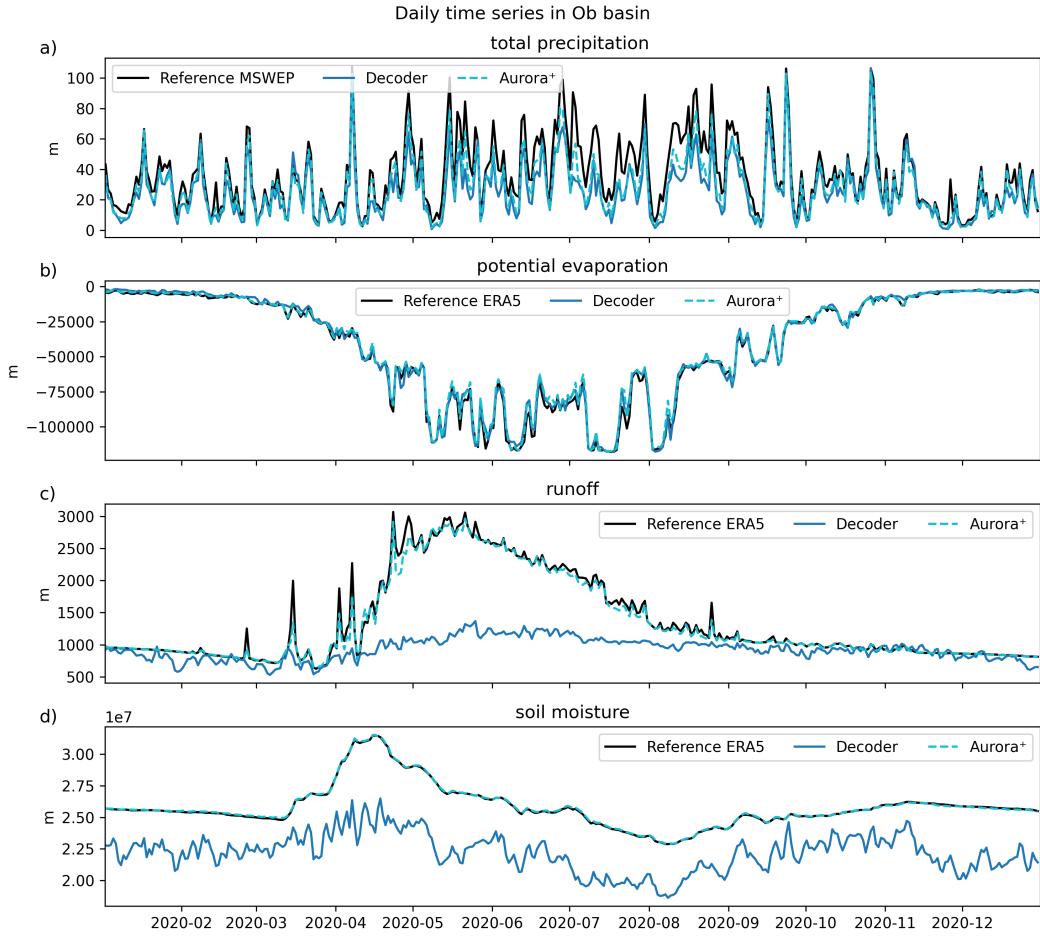


Figure E1: Daily average of 6-hour time series of a) total precipitation [m/6hr], b) potential evaporation [m of equivalent water/6hr], c) runoff [m], d) soil moisture [m] in the Ob basin (shown in green in Fig. 6). Time series are obtained as the sum of pixels inside the basin.

Appendix F Regional Plots

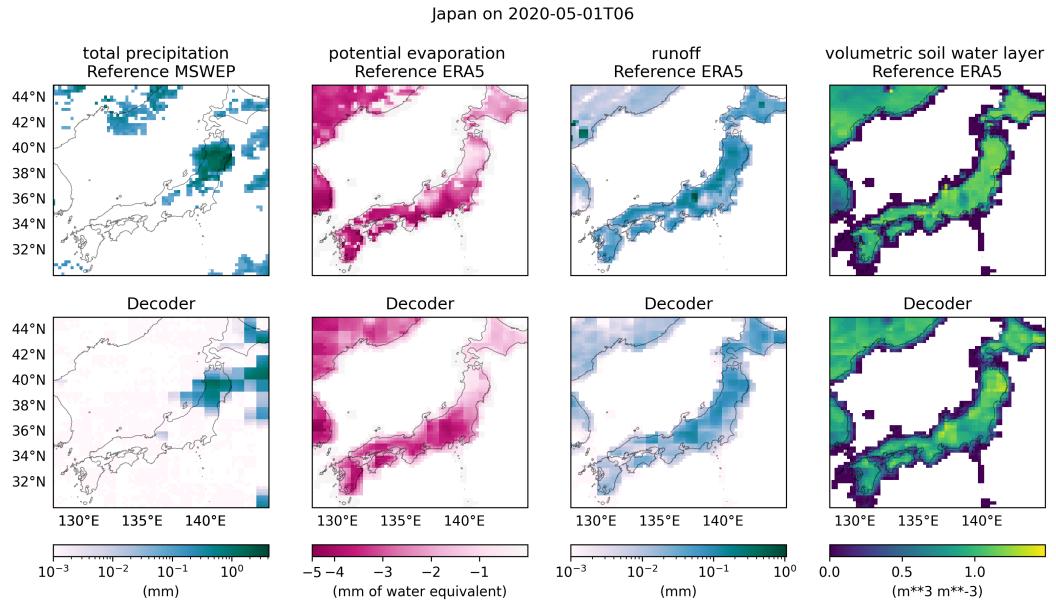


Figure F1: Total precipitation (first column), potential evaporation (second column), runoff (third column), and volumetric soil water layer (fourth column) maps with a zoom around Japan. The decoder prediction (second row) is compared with the reference dataset (first row)

Open Research Section

The ERA5 training data and the test data for the benchmark models (IFS, Graph-Cast, FuXi) were downloaded from the WeatherBench2 Google Cloud bucket <https://console.cloud.google.com/storage/browser/weatherbench2>. The MSWEP precipitation dataset is freely available on <https://www.globo2o.org/mswep/>. The inference code of the decoders is available at <https://github.com/lehmannfa/aurora-lite-decoder>.

Acknowledgments

The authors sincerely thank Dr. Salman Mohebi (SDSC) and Dr. Yun Cheng (SDSC) for their contribution to the implementation of the training pipeline. This publication was made possible by an ETH AI Center postdoctoral fellowship to Fanny Lehmann. This work was supported as part of the Swiss AI Initiative by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID a01 on Alps.

References

- Adamov, S., Oskarsson, J., Denby, L., Landelius, T., Hintz, K., Christiansen, S., ... Schemm, S. (2025). *Building Machine Learning Limited Area Models: Kilometer-Scale Weather Forecasting in Realistic Settings*. arXiv. doi: 10.48550/ARXIV.2504.09340
- Beck, H. E., Wood, E. F., Pan, M., Fisher, C. K., Miralles, D. G., Van Dijk, A. I. J. M., ... Adler, R. F. (2019). MSWEP V2 Global 3-Hourly 0.1° Precipitation: Methodology and Quantitative Assessment. *Bulletin of the American Meteorological Society*, 100(3), 473–500. doi: 10.1175/BAMS-D-17-0138.1
- Bodnar, C., Bruinsma, W. P., Lucic, A., Stanley, M., Allen, A., Brandstetter, J., ... Perdikaris, P. (2025). A foundation model for the Earth system. *Nature*. doi: 10.1038/s41586-025-09005-y
- Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., ... Liang, P. (2021). *On the opportunities and risks of foundation models*. Retrieved from <https://crfm.stanford.edu/assets/report.pdf>
- Chen, L., Zhong, X., Zhang, F., Cheng, Y., Xu, Y., Qi, Y., & Li, H. (2023). FuXi: a cascade machine learning forecasting system for 15-day global weather forecast. *npj Climate and Atmospheric Science*, 6(1), 190. doi: 10.1038/s41612-023-00512-1
- Gebrechorkos, S. H., Leyland, J., Dadson, S. J., Cohen, S., Slater, L., Wortmann, M., ... Darby, S. E. (2024). Global-scale evaluation of precipitation datasets for hydrological modelling. *Hydrology and Earth System Sciences*, 28(14), 3099–3118. doi: 10.5194/hess-28-3099-2024
- Gou, J., & Soja, B. (2024). Global high-resolution total water storage anomalies from self-supervised data assimilation using deep learning algorithms. *Nature Water*, 2(2), 139–150. doi: 10.1038/s44221-024-00194-w
- Hardy, L., & Finney, I. (2025). Leveraging state-of-the-art AI models to forecast wind power generation using deep learning. *Meteorological Applications*, 32(2), e70038. doi: 10.1002/met.70038
- Herde, M., Raonić, B., Rohner, T., Käppeli, R., Molinaro, R., de Bézenac, E., & Mishra, S. (2024, May). Poseidon: Efficient Foundation Models for PDEs. In *Advances in neural information processing systems*. arXiv. Retrieved 2024-06-02, from <https://openreview.net/forum?id=JC1VKK3UXk> (foundation)
- Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., ... Thépaut, J.-N. (2020). The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730), 1999–2049. doi: 10.1002/qj.3803
- Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., & Carreira, J. (2021). Perceiver: General Perception with Iterative Attention. In M. Meila & T. Zhang (Eds.), *Proceedings of the 38th International Conference on*

- Machine Learning* (Vol. 139, pp. 4651–4664). PMLR. Retrieved from <https://proceedings.mlr.press/v139/jaegle21a.html>
- Lam, R., Sanchez-Gonzalez, A., Willson, M., Wirnsberger, P., Fortunato, M., Alet, F., ... Battaglia, P. (2023). Learning skillful medium-range global weather forecasting. *Science*, eadi2336. doi: 10.1126/science.adl2336
- Lehner, B., & Grill, G. (2013). Global river hydrography and network routing: baseline data and new approaches to study the world's large river systems. *Hydrological Processes*, 27(15), 2171–2186. doi: 10.1002/hyp.9740
- Lessig, C., Luise, I., Gong, B., Langguth, M., Stadtler, S., & Schultz, M. (2023). *AtmoRep: A stochastic model of atmosphere dynamics using large scale representation learning*. arXiv. doi: 10.48550/ARXIV.2308.13280
- Li, F., Kusche, J., Sneeuw, N., Siebert, S., Gerdener, H., Wang, Z., ... Tian, K. (2024). Forecasting Next Year's Global Land Water Storage Using GRACE Data. *Geophysical Research Letters*, 51(17), e2024GL109101. doi: 10.1029/2024GL109101
- Liu, C.-C., Hsu, K., Peng, M. S., Chen, D.-S., Chang, P.-L., Hsiao, L.-F., ... Kuo, H.-C. (2024). Evaluation of five global AI models for predicting weather in Eastern Asia and Western Pacific. *npj Climate and Atmospheric Science*, 7(1), 221. doi: 10.1038/s41612-024-00769-0
- Nguyen, T., Brandstetter, J., Kapoor, A., Gupta, J. K., & Grover, A. (2023). Climax: A foundation model for weather and climate. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, & J. Scarlett (Eds.), *Proceedings of the 40th international conference on machine learning* (Vol. 202, pp. 25904–25938). Retrieved from <https://proceedings.mlr.press/v202/nguyen23a.html>
- Palazzoli, I., Ceola, S., & Gentile, P. (2025). GRAiCE: reconstructing terrestrial water storage anomalies with recurrent neural networks. *Scientific Data*, 12(1), 146. doi: 10.1038/s41597-025-04403-3
- Ramdas, A., Garcia, N., & Cuturi, M. (2015). *On Wasserstein Two Sample Testing and Related Families of Nonparametric Tests*. arXiv. doi: 10.48550/ARXIV.1509.02237
- Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., & Thuerey, N. (2020). WeatherBench: A Benchmark Data Set for Data-Driven Weather Forecasting. *Journal of Advances in Modeling Earth Systems*, 12(11), e2020MS002203. doi: 10.1029/2020MS002203
- Rasp, S., Hoyer, S., Merose, A., Langmore, I., Battaglia, P., Russel, T., ... Sha, F. (2024). *WeatherBench 2: A benchmark for the next generation of data-driven global weather models*. arXiv. Retrieved from <http://arxiv.org/abs/2308.15560>
- Roberts, N. M., & Lean, H. W. (2008). Scale-Selective Verification of Rainfall Accumulations from High-Resolution Forecasts of Convective Events. *Monthly Weather Review*, 136(1), 78–97. doi: 10.1175/2007MWR2123.1
- Rodwell, M. J., Richardson, D. S., Hewson, T. D., & Haiden, T. (2010). A new equitable score suitable for verifying precipitation in numerical weather prediction. *Quarterly Journal of the Royal Meteorological Society*, 136(650), 1344–1363. doi: 10.1002/qj.656
- Schmude, J., Roy, S., Trojak, W., Jakubik, J., Civitarese, D. S., Singh, S., ... Ramachandran, R. (2024). *Prithvi WxC: Foundation Model for Weather and Climate*. arXiv. doi: 10.48550/ARXIV.2409.13598
- Wang, X., Liu, S., Tsaris, A., Choi, J.-Y., Aji, A. M., Fan, M., ... Balaprakash, P. (2024). ORBIT: Oak ridge base foundation model for earth system predictability. In *SC24: International conference for high performance computing, networking, storage and analysis* (pp. 1–11). doi: 10.1109/SC41406.2024.00007
- Watt-Meyer, O., Henn, B., McGibbon, J., Clark, S. K., Kwa, A., Perkins, W. A., ... Bretherton, C. S. (2024). *ACE2: Accurately learning subseasonal to decadal atmospheric variability and forced responses*. arXiv. doi:

10.48550/arXiv.2411.11268

Xu, P., Zheng, X., Gao, T., Wang, Y., Yin, J., Zhang, J., ... Chen, X. (2025). An artificial intelligence-based limited area model for forecasting of surface meteorological variables. *Communications Earth & Environment*, 6(1). doi: 10.1038/s43247-025-02347-5

Zhao, X., Zhou, Z., Zhang, W., Liu, Y., Chen, X., Gong, J., ... Bai, L. (2024). *WeatherGFM: Learning A Weather Generalist Foundation Model via In-context Learning*. arXiv. doi: 10.48550/ARXIV.2411.05420