

# Multi-fidelity climate model parameterization for better generalization and extrapolation

Mohamed Aziz Bhouri<sup>1,\*</sup>, Liran Peng<sup>2</sup>, Michael S. Pritchard<sup>2,3</sup>, and Pierre Gentine<sup>1,4</sup>

<sup>1</sup>Columbia University, Earth and Environmental Engineering, New York City, 10027, USA

<sup>2</sup>University of California, Irvine, Center for Complex Biological Systems, Irvine, 92697, USA

<sup>3</sup>NVIDIA, Santa Clara, CA, 95051, USA

<sup>4</sup>Columbia University, Climate School, New York City, 10027, USA

\*mb4957@columbia.edu

## ABSTRACT

Machine-learning-based parameterizations (i.e. representation of sub-grid processes) of global climate models or turbulent simulations have recently been proposed as a powerful alternative to physical, but empirical, representations, offering a lower computational cost and higher accuracy. Yet, those approaches still suffer from a lack of generalization and extrapolation beyond the training data, which is however critical to projecting climate change or unobserved regimes of turbulence. Here we show that a multi-fidelity approach, which integrates datasets of different accuracy and abundance, can provide the best of both worlds: the capacity to extrapolate leveraging the physically-based parameterization and a higher accuracy using the machine-learning-based parameterizations. In an application to climate modeling, the multi-fidelity framework yields more accurate climate projections without requiring major increase in computational resources. Our multi-fidelity randomized prior networks (MF-RPNs) combine physical parameterization data as low-fidelity and storm-resolving historical run's data as high-fidelity. To extrapolate beyond the training data, the MF-RPNs are tested on high-fidelity warming scenarios, +4K, data. We show the MF-RPN's capacity to return much more skillful predictions compared to either low- or high-fidelity (historical data) simulations trained only on one regime while providing trustworthy uncertainty quantification across a wide range of scenarios. Our approach paves the way for the use of machine-learning based methods that can optimally leverage historical observations or high-fidelity simulations and extrapolate to unseen regimes such as climate change.

## Introduction

Due to limited computational resources and the many scales required in climate or turbulent simulations, unresolved sub-grid processes are approximated through parameterization schemes, or closures in numerical models. Parameterizations serve as approximate representations of small-scale processes and are the most dominant source of uncertainty in models predictions. To reduce these structural closures errors and uncertainties, several recent pieces of work have proposed machine-learning based parameterizations, which have been shown to dramatically improve the representation of physical processes and strongly reduce structural errors compared to standard schemes<sup>1–11</sup>. Another source of uncertainty stems from the inherent stochastic nature of many physical sub-grid processes in nature, such as turbulence or cloud micro-physics<sup>12–15</sup>. Stochastic parameterization schemes have been proposed to better characterize this latter source of uncertainty, as it can be important to correctly predict the prediction variability<sup>16–25</sup>.

Along with the development of recent climate parameterization schemes, various climate simulation data have been made available. However, most of these were built with simple aquaplanets<sup>2,3,26–29</sup> and those that considered real geography<sup>30,31</sup> did not include enough variables for a complete land-surface coupling. Hence, there is a wealth of relatively low-fidelity climate simulation data that is available to build climate pa-

rameterization schemes, while high-fidelity datasets based on high-resolution and/or multi-scale climate simulations are rare. Therefore, there is a clear need to investigate the possibility of building schemes that take advantage of the abundant low-fidelity data in order to improve high-fidelity parameterizations. In addition, although several machine learning-based methods have been successfully developed in order to parameterize turbulence<sup>32</sup>, atmospheric<sup>1–4,30,33–35</sup> and oceanic processes<sup>5</sup>, these methods struggle with out-of sample testing inputs and are unable to extrapolate beyond the training data regimes and scenarios<sup>2</sup> (out-of-distribution limitations). An important body of recent work has made exciting progress on using machine learning methods to reduce biases in climate simulations<sup>36–39</sup>. However, these approaches were restricted to improving coarse-grid climate and weather models using higher resolution simulations while the opposite would of greater use given the abundant low-fidelity data.

Multi-fidelity (MF) models have recently been successful in several computational science and engineering applications<sup>40–46</sup>. These models are suitable for problems where multiple datasets or computational models are available for a given system of interest. MF models aggregate data and information with different fidelity, i.e. level of accuracy and details availability<sup>47</sup>. High-fidelity (HF) models or datasets provide more accurate information but require greater computational or measurement resources. On the other hand, low-fidelity

(LF) models or datasets are less accurate but cheaper to run or obtain, and hence generally more abundant compared to HF simulation runs or data<sup>48</sup>.

In this work we use a probabilistic MF approach in order to allow uncertainty quantification. Different Bayesian models can be used in order to build MF approaches including: Markov-Chain Monte Carlo (MCMC) sampling methods<sup>49</sup>, variational inference techniques<sup>50,51</sup>, deep ensembles<sup>52,53</sup> and dropout<sup>54,55</sup>. Given the typical dimensionality and size of the datasets for Earth System Model (ESM) parameterizations, the gold-standard MCMC methods are out of scope. Besides, variational inference approximations can suffer from posterior variance underestimation and results in a poor approximation of the true multi-modal posterior distribution when applied to deep learning frameworks<sup>56</sup>. In addition, it has been shown that dropout and standard deep ensemble methods often provide minimal uncertainty estimates which prevents their use in applications requiring sufficiently accurate approximation of posterior distributions<sup>56,57</sup>.

Randomized Prior Networks (RPNs)<sup>58</sup> were developed in order to provide a compromise between acceptable computational cost for building Bayesian surrogate models and overcome the uncertainty underestimation. RPNs take advantage of an explicit incorporation of prior knowledge in order to improve the model predictions in regions where limited or no training data is available<sup>58–60</sup>. There has been additional theoretical studies proving the conservative uncertainty obtained with RPNs and their ability to reliably detect out-of-distribution samples<sup>61</sup>. RPNs have also been proven to outperform HMC methods, variational inference techniques and dropout as a Bayesian approximation in the context of complex sequential decision making tasks<sup>56,61</sup>. The RPNs improvement is mainly driven by their parallelizable implementation resulting in a significantly lower computational cost and the possibility of building Bayesian surrogate models for complex and large neural network architectures.

Extending on previous deterministic neural network parameterization studies of atmospheric ESM parameterization<sup>30</sup>, here we propose a multi-fidelity RPN model (MF-RPN) as a parameterization scheme for atmospheric convection (deep clouds), which is the first of its kind to the best of our knowledge. The MF-RPN surrogate model is designed to take into account the distribution shift across regimes leveraging the rich LF training data regimes while refining it with higher accuracy but more limited HF regimes. This proves crucial to obtain skillful extrapolation predictions for unseen HF testing data. We show that the proposed MF-RPN can provide the best of both worlds: the higher accuracy of the HF data and the generalization capability thanks to the LF one. The improved MF-RPN skillful predictions are tested across various error metrics on HF data of unseen warmer climate scenarios and against three other surrogate models.

## Results

### Problem setup

We define the convection superparameterization problem considered and the used ESM datasets.

### *ESM convection superparameterization*

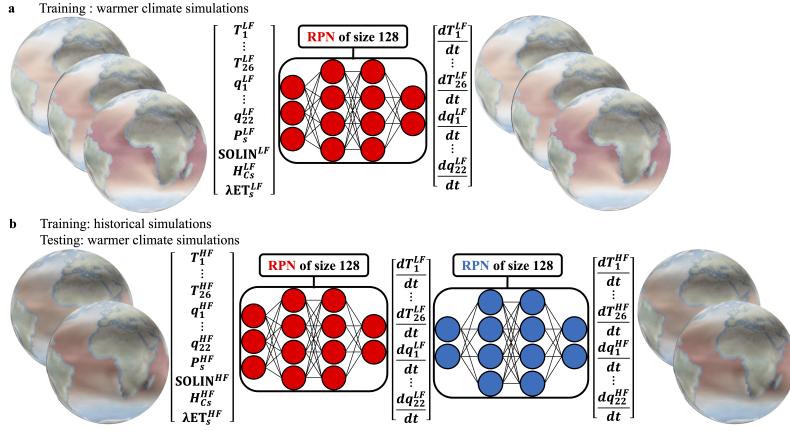
The problem considered here consists of predicting subgrid-scale tendencies of heat and moisture convection (i.e. time rate of change) at all vertical levels and for every timestep<sup>30</sup>. The convection parameterization for climate models is becoming a mature problem given the recent studies focusing on it<sup>2,3,26,62,63</sup>. The parameterization input is similar to the standard Community Earth System Model version 2.1.3 (CESM2.1.3) Community Atmospheric Model version 5 (CAM5) parameterization and is taken as coarse-grid atmospheric thermodynamics components consisting of: atmospheric temperature for each of the 26 vertical levels spanning the column and specific humidity for each of the 22 vertical levels spanning the column except the first four levels from top of atmosphere (TOA) (Methods). The input vector also contains the surface pressure, TOA solar insulation, surface latent heat flux and surface sensible heat flux (figure 1).

The parameterization output is the subgrid-scale convective tendency of temperature, or heat tendency for short, and the subgrid-scale convective tendency of specific humidity throughout the column, or moisture tendency for short (figure 1). This tendency definition accounts for the sub-grid advection of temperature and moisture by convection and fine-scale turbulence, as well as for the effect of radiative heating throughout the column on temperature tendency.

### Datasets

Our evaluation consists of the CESM (high-fidelity) Superparameterized Community Atmospheric Model version 5 (SPCAM5) in a real geography setup (Methods). Unlike CAM5, SPCAM5 nearly explicitly resolves atmospheric moist convection (including deep convection) by using idealized embedded cloud resolving models<sup>64,65</sup>, introducing less physical approximations. Both models incorporate the CAM radiation package (CAM-RT)<sup>66,67</sup>. The HF training data, of size 29.5 M points (pixels x times), is constructed by considering the SPCAM5 historical run simulation (i.e. non-global warming) of three months (Methods).

The proposed multi-fidelity model aggregates high-fidelity SPCAM5 historical dataset and low-fidelity (historical and future) CAM5 data. SPCAM5 has a much higher computational cost and better capability to resolve convection compared to CAM5. CAM5 uses a physically-based parameterization of convection, introducing more physical approximations compared to SPCAM5, which resolves deep convection. In addition, radiation in CAM5 is estimated every 2 time-steps, while the radiation is estimated every 1 time-step for SPCAM5 by default. Therefore, CAM5 is a good candidate for a low-fidelity model of SPCAM5, as it is computationally cheaper but also less accurate. For all CAM5 and SPCAM5 simulations, the cosine of the solar zenith angle is estimated as a



**Figure 1. Multi-fidelity problem setting for ESM parameterization.**  $T_i, i = 1, \dots, 26$  refer to atmospheric temperature [K] at different vertical levels.  $q_i, i = 1, \dots, 22$  refer to specific humidity [kg/kg] at different vertical levels.  $P_s$ , SOLIN,  $H_{Cs}$  and  $\lambda ET_s$  refer to surface pressure [Pa], TOA solar insulation [ $W/m^2$ ], sensible heat flux [ $W/m^2$ ] and latent heat flux [ $W/m^2$ ] respectively. Parameterization input and output are of dimension 52 and 48 respectively. **a**, Low-fidelity problem setting for multi-fidelity RPN-based CAM5/SPCAM5 convection superparameterization. **b**, High-fidelity problem setting for multi-fidelity RPN-based CAM5/SPCAM5 convection superparameterization

function of Julian calendar day, latitude, longitude and Solar declination.

Since we are interested in warmer climate scenarios, the CAM5 simulation corresponding to a global warming situation, with a prescribed sea surface temperature (SST) that has been augmented by 4 K and 8 K, referred to as +4K and +8K simulations respectively, were considered as LF data candidates for training. A comparison of their inputs and outputs' distributions with those of the SPCAM5 training data shows a more pronounced extrapolation regime for CAM5 +8K data, mainly due to the increased holding capacity of moisture in the atmosphere with climate change (Clausius-Clapeyron) at higher temperatures. Hence, the CAM5 +8K was chosen as low-fidelity model for extrapolation (Methods).

Since we are interested in extrapolating beyond the training data, the test data is constructed by considering the CESM SPCAM5 +4K simulation. In order to enhance the extrapolation evaluation to unseen data, the testing dataset corresponds to a full year of the +4K simulation, resulting in a final test dataset of roughly 121.1 M points (Methods). Beyond global warming, the test dataset also extrapolates to unseen phases of the SPCAM5 seasonal cycle.

## Surrogat models

We provide description of the four surrogate models that are built for the CAM5/SPCAM5 convection superparameterization.

### Single-fidelity Randomized Prior Networks

In this work, Bayesian models are constructed using an ensemble method called Randomized Prior Networks (RPNs)<sup>58</sup>. Each member of the RPNs is built as the sum of a trainable and a non-trainable (so-called “prior”) surrogate model; we used fully-connected neural network for simplicity. Multiple

replicas of the networks are constructed by independent and random sampling of both trainable and non-trainable parameters<sup>60,68</sup>. The non-trainable parameters are initialized but then kept fixed during the fitting process which only optimizes over the trainable parameters. In our case of fully-connected neural networks, we resort to Glorot initialization<sup>69</sup>, which defines the probability distributions from which the fixed non-trainable parameters are sampled. RPNs also resort to data bootstrapping in order to mitigate a potential uncertainty collapse of the ensemble method when tested beyond the training data points<sup>60</sup>. Data bootstrapping consists of sub-sampling and randomization of the data on which each network in the ensemble is trained.

The Single High Fidelity model corresponds to a standard RPN trained only on the HF data and will be referred to as SF-HF-RPN. Hyperparameters of individual neural networks did not need to be tuned from scratch. They were instead chosen based on the hyperparameter optimization over  $\sim 250$  trials conducted in *Mooers et al.*'s study on fully-connected neural network convection superparameterization for SPCAM5<sup>30</sup> (Methods). RPN ensembles of 128 networks were considered as justified in *Yang et al.*<sup>68</sup>.

### Deterministic neural network

In addition to the SF-HF model, we also considered, for reference, a deterministic model defined as a single fully-connected neural network with the same hyperparameters as the SF-HF model's individual neural networks. Both deterministic and SF-HF-RPN models were trained on the SPCAM5 HF historical run data providing a baseline for high-fidelity models trained only on historical data.

### **Multi-fidelity Randomized Prior Networks**

The multi-fidelity model is also constructed using RPNs of size 128. Trainable and non-trainable surrogate models of each member of the multi-fidelity RPN (MF-RPN) are built with the architecture detailed in figure 1.b. The chosen architecture consists of two fully connected deep neural networks. The first network (highlighted in red in figure 1) predicts the low-fidelity parameterization output from the parameterization input, while the second network (highlighted in blue in figure 1) predicts the high-fidelity parameterization output as a function of the low-fidelity parameterization output. The trainable surrogate model of each member of the MF-RPN is trained using a joint training of both networks (Methods).

Our MF-RPN learns the mappings between related physical variables: emulating the parameterization (inputs to outputs) at low-fidelity (red network in figure 1), and mapping the parameterization outputs at different fidelity levels (low to high-fidelity, blue network in figure 1.b). The proposed architecture directly learns the non-linear mapping between the low-to-high fidelity outputs instead of inferring the difference between them as an error bias correction<sup>37, 38, 70, 71</sup>. The bias correction approach was only shown to improve coarse-grid climate models using higher resolution simulations and not vice versa despite the abundant low-fidelity data. In addition, the chosen architecture naturally accommodates outputs of different dimensions for different fidelity levels. In the case of inputs of different dimensions for different fidelity levels, an additional neural network can be added in order to infer the mapping between the different inputs. Besides, the chosen architecture naturally ensures uncertainty propagation between different fidelity levels since low-fidelity predictions are directly fed as inputs for the high-fidelity model within the MF-RPN (Methods). Finally, since the low-fidelity training data was built such that it provides the MF model with useful information regarding the high-fidelity extrapolation scenarios, the MF-RPN model is trained on normalized data with respect to the statistics of the CAM5 +8K run data in order to take into account the data distribution shift between different fidelity levels (Methods).

### **Low-fidelity Randomized Prior Networks**

A low-fidelity RPN model can be considered based on the MF-RPN model detailed above without any further training. Indeed, the low-fidelity network within the MF-RPN model (red network in figure 1) already provides predictions for the convection parameterization outputs. Hence we can also test this LF-RPN model on high-fidelity data points by considering the corresponding parameterization inputs. The LF-RPN can be seen as a control model whose performance allows assessing whether the MF-RPN model is capable of properly aggregating both datasets to well generalize beyond the training data. If both models performance are similar, then the MF-RPN improvement would solely be due to being trained on the abundant low-fidelity data for a warmer climate and with a full seasonal cycle. However, if the MF-RPN results improves upon the LF-RPN ones, then it would justify that

the MF-RPN model is well capable of merging both training datasets, including the scarce but more physically sound high-fidelity data even without the full seasonal cycle.

### **Forecast skills**

All surrogate models are evaluated based on their performance on the high-fidelity test dataset corresponding to the SPCAMS +4K simulation.

### **Evaluation metrics**

Different evaluation metrics are considered and computed for each output variable. We report the mean absolute error (MAE) and the coefficient of determination ( $R^2$ ). The MAE is always positive and a lower value corresponds to a more accurate model. The coefficient of determination is upper-bounded by 1 and values closer to 1 correspond to more accurate models (Methods).

### **Forecast skills results**

For the heat tendency, the MF-RPN is the only model with positive global  $R^2$  values for all vertical levels, with an average  $R^2$  of 0.62 across all levels (figure 2.a and figure 6.a in Supplementary Information showing the negative values for  $R^2$  where appropriate). Besides, MF-RPN is always the best model except for the the 137 and 160hPa vertical levels. Except for the lowest and highest vertical levels, LF-RPN is outperforming both deterministic NN and SF-HF-RPN models (figure 2.a). Hence, for these two levels, historical SPCAMS simulations are closer to those of SPCAM5 +4K run. However, for any other vertical level except the first one at TOA and the closest one to sea surface, CAM5 +8K simulation provides a better approximate of SPCAM5 +4K run dynamics. For most of the vertical levels beyond the two extreme ones, MF-RPN is improving upon LF-RPN which in turn is outperforming deterministic NN and SF-HF-RPN models. In addition, for lowest and highest vertical levels, MF-RPN is improving upon the deterministic NN and SF-HF-RPN models which in turn are outperforming the LF-RPN model. Hence, the MF-RPN has the ability to get the best of the both worlds by aggregating both datasets of different fidelity levels as (1) it learns from a high-fidelity parameterization that resolves convection based on the high-fidelity dataset and (2) generalizes beyond the high-fidelity training data regime thanks to the informative low-fidelity simulations covering regimes at higher sea surface temperatures. It is worth noting that the SF-HF-RPN is capable of improving upon the deterministic NN for nearly all vertical levels and even for deterministic error metrics (figure 2.a), showing the benefits of using RPNs as a stochasticity-aware surrogate model.

For the moisture tendency, the overall performance of all models for almost all vertical levels is lower in terms of  $R^2$  compared to the results obtained for the heat tendency (figure 2). This result can be mainly attributed to the higher stochasticity of humidity and precipitation compared to the temperature. The MF-RPN model is the best performing model for all pressure levels except for the 188hPa vertical

level where the LF-RPN is the best one, and for the closest level to the surface (958hPa) where MF-RPN is outperformed by the deterministic NN and SF-HF-RPN (figure 2.b). It is worth noting that for this level, MF-RPN is still performing well with an  $R^2 = 0.71$ , unlike LF-RPN showing a negative  $R^2 = -0.6$  (figure 6.b). For all pressure levels where moisture tendency is the most significant and critical for cloud formation (typically between 250 and 750hPa), the MF-RPN model clearly outperforms all other models with an average  $R^2$  equal to 0.73 across different vertical levels (figure 2.b). In addition, for all levels where the deterministic NN, SF-HF-RPN and LF-RPN all fail (e.g. all levels below 160hPa, 897 and 937hPa), the MF-RPN is still capable of providing significantly better results than all these models showing even positive  $R^2$  values (e.g. 0.36 and 0.32 for levels 897 and 937hPa) thanks to both datasets aggregation.

The LF-RPN is outperforming the deterministic NN. and SF-HF-RPN models except within the stratosphere (where convection is absent anyways) and for pressure levels close to the surface (figure 2.b). This result confirms that within the highest and lowest vertical levels, historical SPCAM5 simulation dynamics are closer to those of SPCAM5 +4K run, while beyond them CAM5 +8K simulation provides a better approximate of SPCAM5 +4K run dynamics. Finally, for most of vertical levels from TOA to 494hPa, the deterministic NN. is outperforming the SF-HF-RPN, while the opposite is observed for all vertical levels from 581 to 958hPa. Hence, the SF-HF-RPN is only capable of better resolving the moisture convection stochasticity for vertical levels below the 494hPa one, while it struggles to do so at higher levels.

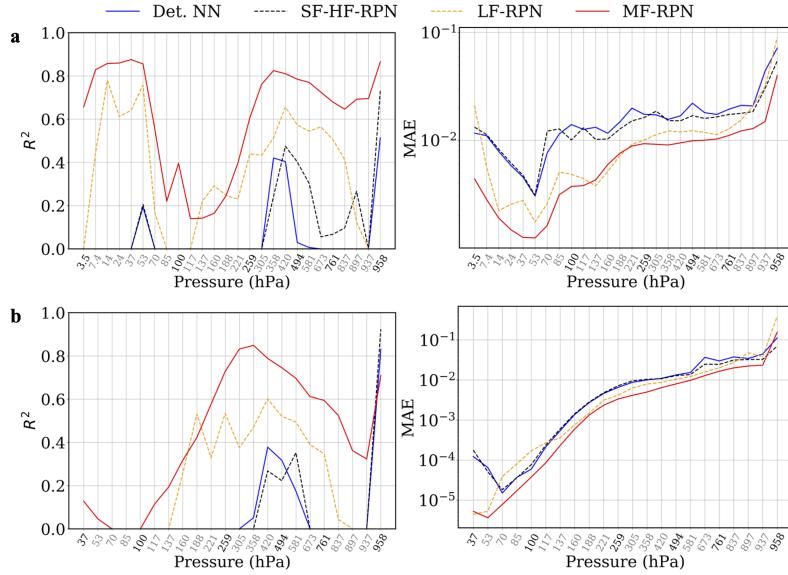
The SF-HF-RPN model has higher  $R^2$  values for the moisture tendency than the deterministic NN in the temperate zone thanks to its capacity of better resolving the moisture convection stochasticity within this region (figure 3 and figure in Supplementary Information). However, the SF-HF-RPN model fails to provide more accurate predictions within the tropics and polar regions. The LF-RPN model improves further upon the SF-HF-RPN model within the temperate zone and even within the tropics and polar regions. These results confirm the informative capacity of the low-fidelity data for the extrapolation scenario of interest and also the LF-RPN capacity to resolve the moisture convection stochasticity since it is an ensemble method. Finally, the MF-RPN model improves even further upon the LF-RPN model across all regions with a nearly perfect  $R^2$  score in the temperate zone. The MF-RPN model also shows better results for all tropical regions (figure 3 and figure in Supplementary Information). This result is of a significant importance since we are extrapolating to warmer climates and hence the tropics (the warmest region of the world) provide test data-points that are well outside the training datasets distributions. In addition, the tropics is a challenging region to model in terms of convection and ESMs exhibit many typical problems within this region that are related to sub-grid convection parameterizations. Among these problems we can mention the double inter-tropical conver-

gence zone (ITCZ)<sup>72</sup>, too much drizzle and missing precipitation extremes<sup>73</sup>, and an unrealistic equatorial wave spectrum with a missing Madden–Julian oscillation (MJO)<sup>74</sup>. Therefore, providing a framework to improve convection parameterization within this region can help remedy these issues.

In Supplementary Information, we provide all longitude-latitude variations of the MAE and  $R^2$  metrics for heat and moisture tendencies for pressure levels: 259, 494 and 761 hPa (figures 7 and 8). These results show a very similar behavior as observed above for the moisture tendency at level 494 hPa. For other levels and for heat tendency, the MF-RPN model shows even higher  $R^2$  values in the south Atlantic ocean and African Sahara. We also provide the temporal variation of the MAE and  $R^2$  metrics in Supplementary Information.

For both tendencies and nearly all vertical levels, the MF-RPN model shows improved results compared to all other surrogate models including for the tropics across all vertical levels between around 250 and 800 hPa, where lies the double ITCZ region (figures 4.a and 4.b). The MF-RPN shows significant improvement for the heat tendency parameterization in the stratosphere, mostly within the polar region and the temperate zone. It also displays a better parameterization for the heat tendency within the first vertical level close to sea surface, showing a better parameterization for boundary layer regions (figure 4.a). For the heat tendency, the SF-HF-RPN model improves upon the deterministic NN. mostly within the tropics thanks to a better stochasticity representation (figure 4.a). However, the improvement is only noticeable between the 250 hPa and 750 hPa pressure levels, which are the critical levels for cloud formation. The LF-RPN model improves further compared to the SF-HF-RPN model for the tropics between the 250 hPa and 750 hPa pressure levels, and also shows higher  $R^2$  values for both polar regions, including a very pronounced improvement in these regions within the stratosphere (figure 4.a). Compared to the LF-RPN, the MF-RPN model improves the heat tendency parameterization results for the south pole across nearly all pressure levels, while it under-performs in the north pole for vertical levels below 300 hPa.

For the moisture tendency, the SF-HF-RPN still shows some improvement compared to the deterministic NN model, mostly within the southern temperate zone (figure 4.b). The LF-RPN model improves further compared to the SF-HF-RPN model for the tropics between the 400 and 600 hPa pressure levels, which is a smaller region compared to the improvement observed for the heat tendency. The LF-RPN model also shows higher  $R^2$  values for both polar regions. The MF-RPN model improves further upon the LF-RPN in the tropical region between the 200 and 900 hPa vertical levels, and also in the temperate zone between 250 and 700 hPa levels for the south hemisphere (figure 4.b). The temperate zone improvement applies to a smaller region mostly located between 250 and 550 hPa levels for the north hemisphere. This observation is coherent with the heat tendency results showing a higher MF-RPN's performance for the temperate



**Figure 2.**  $R^2$  and MAE metrics evaluation for different models across all test data points concatenated over space and time. Negative  $R^2$  values are lumped to 0 for clarity purposes. **a**, Heat tendency results. **b**, Moisture tendency results

zone in the southern hemisphere compared to the northern one.

Finally, we verify that the MF-RPN's uncertainty quantification estimated over the ensemble predictions is coherent as it increases with the predictions error (figure 12 in Supplementary Information). This means that without accessing any information on the true target values, the MF-RPN model is intrinsically capable of estimating its predictions accuracy across different test data points. We also verify that the longitude-latitude structure of the uncertainty well matches with the longitude-latitude variation of the predictions error, with the highest values being observed around the tropics where the inherent stochasticity of convection is the highest compared to other regions (figure 13).

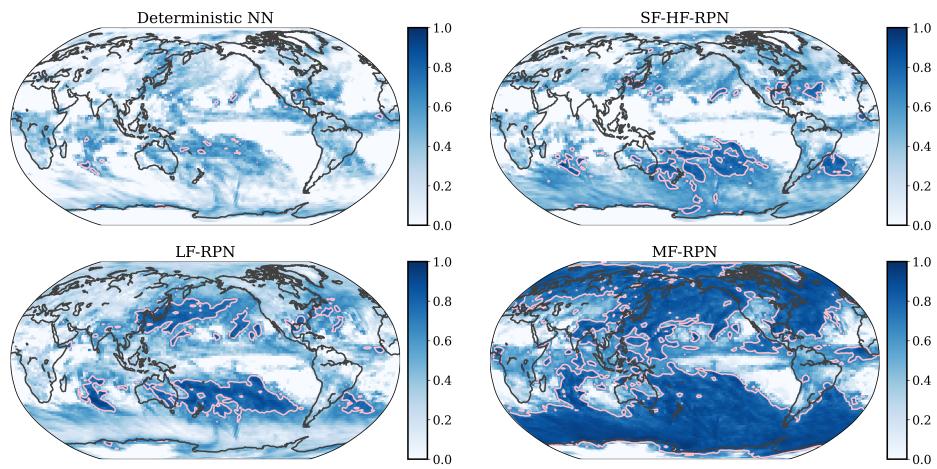
## Discussion

Extrapolation beyond training datasets is a long-standing problem of importance for machine-learning-based models, and for the emulation of physical models in particular. In this work, we showed how the proposed multifidelity (here with an RPN) approach can tackle this problem by considering the high-fidelity convection data on historical observations and optimally merging it with a prior coming from a physically based parameterization exploring more diverse regimes as it is computationally cheaper. We showed that the proposed approach can extrapolate heat and moisture convection predictions over substantial climate warming situations, where existing supervised (single-fidelity) methods struggle. The improvement includes even the tropics where convection stochasticity is higher compared to other regions and where different Earth system models exhibit many typical problems related to sub-grid convection parameterizations. We also verified

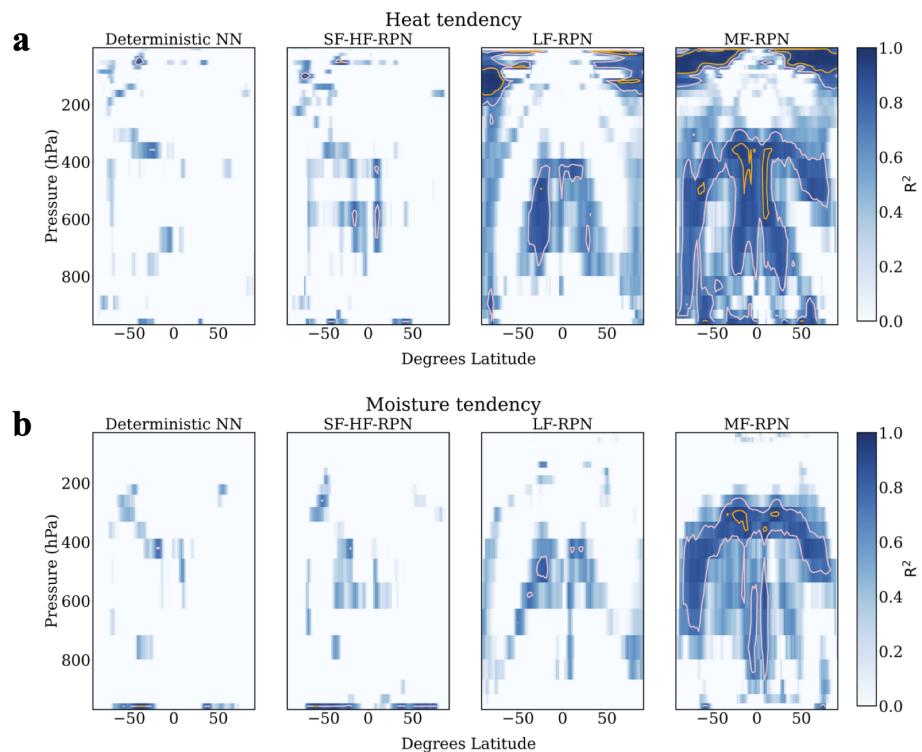
that the proposed multifidelity-RPN uncertainty quantification coherently increases with predictions error. The proposed MF parameterization approach can also be combined with explainable AI techniques to further study similarities and discrepancies between different Earth system models.

The multifidelity-RPN performance is due to the model's design (architecture accounting for data distribution shift) and to its optimal aggregation of different datasets of different fidelity levels. This latter property allows the MF model to provide the best of both worlds: the capacity to extrapolate based on low-fidelity data exploring many regimes of convection and the higher accuracy based on high-fidelity data covering more limited regimes because of its computational cost. Hence, the MF-based parameterization narrows further the gap between the climate science and machine-learning communities by (1) building trust in the capacity of ML-based parameterization to extrapolate to unknown scenarios thanks to its low-fidelity component, (2) while also harnessing more physically-consistent and higher accuracy high-fidelity data.

There is still room for improving the proposed multifidelity parameterization scheme by enforcing physical constraints<sup>75,76</sup>, aggregating observational data and extending it to an online setting within differentiable solvers when available<sup>32,77</sup>. Nonetheless, whereas existing machine learning-based climate parameterizations struggle to generalize beyond the training data regimes, we hope that thanks to the multifidelity extrapolation capabilities, this work will pave the way to finally tackle climate change projection with Artificial Intelligence.



**Figure 3.** Longitude-latitude variation of coefficient of determination  $R^2$  for moisture tendency at vertical level  $P = 494$  hPa for different surrogate models.  $R^2$  is evaluated on the test dataset and negative values are lumped to 0 for clarity purposes.



**Figure 4.** Pressure-latitude variation of coefficient of determination  $R^2$  for different surrogate models.  $R^2$  is evaluated on the test dataset and negative values are lumped to 0 for clarity purposes. **a**, Heat tendency results. **b**, Moisture tendency results.

## Methods

### Earth system model convection superparameterization

Our base model is the CESM2.1.3 CAM5 model with real geography boundary conditions. CAM5 uses a physically-based parameterization of convection and is hence taken as low-fidelity model. The high-fidelity model is taken as the super-parameterized CAM version 5 model (SPCAM5). Notably, while CAM5 employs certain standard packages, SPCAM5 distinguishes itself by its capability to explicitly resolve sub-grid scale physical processes, making it computationally broader in scope<sup>78</sup>. Within CAM5, the micro-physics is driven by the two-moment bulk strati-form cloud micro-physics scheme<sup>79</sup>. CAM5 macro-physics draws from Park and Bretherton's shallow convection and moist turbulence schemes<sup>80</sup>, and its planetary boundary layer (PBL) packages are based on Bretherton and Park' moist turbulence parameterization<sup>81</sup>. SPCAM5 uses idealized cloud resolving models (CRM) in order to nearly explicitly resolve atmospheric moist convection. In particular, SPCAM5 uses the one-moment cloud micro-physics. The SPCAM5 runs considered use 32 CRM columns and 25 CRM vertical levels.

For SPCAM5 training and testing datasets considered, the first 4 vertical levels starting form Top Of the Atmosphere (TOA) show all zero values for the moisture tendency across all earth and for the whole simulations time periods. Therefore, the first 4 vertical levels starting form TOA have been discarded for the moisture tendency in the parameterization problem, and coherently for the specific humidity.

### Datasets

All CAM5 and SPCAM5 simulations considered commence using climatological input data derived from a 20-year mean span around the year 2000. This data includes relevant solar radiation, greenhouse gas levels, oxidant concentrations, and present-day aerosol emissions (denoted as F2000). The prescribed SST and sea ice data sets were constructed as a blended product, using the global HadISST OI data set<sup>82</sup>. The considered forcing consists of annually repeating climatological SSTs with full seasonality. In the simulations labeled +4K and +8K, the standard SST is elevated by 4K and 8K respectively.

The high-fidelity SPCAM5 training data is constructed by considering a historical run simulation while allowing for a model spin-up of a month. The training data corresponds to the time period from February 1st 2003 to April 31st 2003. The horizontal grid resolution of the ESM consists of a  $1.9^\circ \times 2.5^\circ$  finite-volume dynamical core (i.e., 13824 grid cells with 96 in latitude and 144 in longitude). The vertical resolution varies from  $\approx 150$  m to  $\approx 5300$  m. The ESM time step is 30 min and a temporal sub-sampling by a factor of 2 is performed (to reduce the overly correlated training data), resulting in a final training dataset of roughly 29.5 M points.

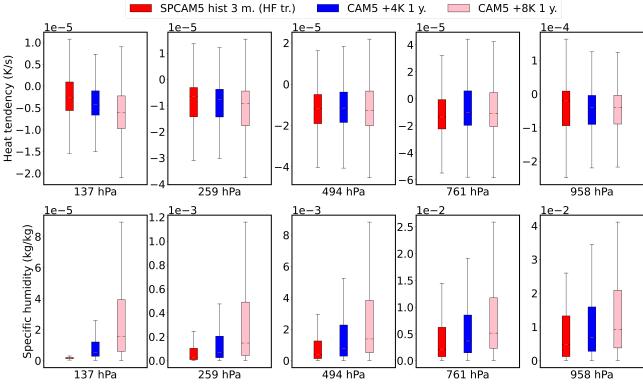
For the high-fidelity SPCAM5 testing data, the corresponding temporal and spatial resolutions, model spin-up and tem-

poral sub-sampling are the same as detailed above for the historical run. The testing dataset corresponds to a full year of the +4K simulation, covering the time period from February 1st 2003 to January 31st 2004, resulting in a final test dataset of roughly 121.1 M points. The testing dataset is constructed with a full-year simulation in order to have a comprehensive analysis of the models performance when tested on unseen climate scenarios and extrapolated to other phases of the SPCAM5 seasonal cycle.

Given the testing dataset defined above, a straightforward choice for the low-fidelity CAM5 training dataset would be to consider a +4K simulation as defined for SPCAM5 for testing. However, low-fidelity data should not be defined with the assumption of prior knowledge of the testing data, but rather on the exploration of scenarios and regimes it provides beyond those observed within the high-fidelity training data. Hence, both CAM5 +4K and +8K are considered as potential candidates. The corresponding temporal and spatial resolutions, model spin-up and temporal sub-sampling are the same as detailed above for SPCAM5 simulations. In the context of multi-fidelity modelling and given the lower CAM5 computational cost, the simulation time period was taken from February 1st 2003 to January 31st 2004 in each simulation. An analysis of the inputs and outputs' distributions for CAM5 +4K and +8K training datasets shows a broader distribution for the CAM5 +8K specific humidity across all pressure levels considered compared to the CAM5 +4K dataset (figure 5). Hence CAM5 +8K provides a broader extrapolation regime due to the increased holding capacity of moisture in the atmosphere with climate change (Clausius-Clapeyron). CAM5 +8K also provides a clearer extrapolation for the heating tendency than CAM5 +4K when compared to the high-fidelity SPCAM5 historical run simulation (figure 5). Based on the data distribution comparison, the CAM5 +8K is chosen as low-fidelity model since it provides a significantly more pronounced extrapolation beyond the regimes spanned by the SPCAM5 training dataset compared to CAM5 +4K. This property proves being crucial in obtaining skillful extrapolation predictions when the multi-fidelity model is tested on unseen SPCAM5 +4K data.

### Multi-fidelity Randomized Prior Networks

The trainable surrogate model of each member  $j$  of the MF-RPN is fitted using a joint training strategy of both networks. Hence the corresponding loss function that is minimized by stochastic gradient descent contains two terms. One term ensures that the first network learns the low-fidelity parameterization (red network in figure 1.a). The second term ensures that the pipeline through both networks learns the high-fidelity parameterization (red and blue networks in figure 1.b). Let  $f_{\theta_{LF},j}$  denote the red network learning the LF parameterization, and  $f_{\theta_{HF},j}$  the blue one learning the mapping to the HF parameterization output. These two networks are trained jointly



**Figure 5.** Data distribution of the specific humidity and heat tendency for SPCAM5 training data (historical simulation) and two potential CAM5 training datasets (+4K and +8K simulations) at 5 different vertical levels.

via the minimization of the following loss function:

$$\mathcal{L} = \frac{1}{N_L} \sum_{i=1}^{N_L} \left( y_{LF,i} - f_{\theta_{LF,j}}(x_{LF,i}) \right)^2 + \frac{1}{N_H} \sum_{i=1}^{N_H} \left( y_{HF,i} - f_{\theta_{HF,j}}(f_{\theta_{LF,j}}(x_{HF,i})) \right)^2, \quad (1)$$

where  $N_L$  and  $N_H$  correspond to the low- and high-fidelity batch sizes respectively. Our MF-RPN learns the mappings between related physical variables: emulating the parameterization (inputs to outputs) at low fidelity via the network  $f_{\theta_{LF,j}}$ , and mapping both parameterization outputs at different fidelity levels (low- to high-fidelity) using the network  $f_{\theta_{HF,j}}$ .

We opted for a joint training of the low- and high-fidelity networks since a sequential training would put more weight and importance on the second network  $f_{\theta_{HF,j}}$  (blue network in figure 1.b) as it will be trained after the fit of network  $f_{\theta_{LF,j}}$  (red network in figure 1.a), which is then held fixed. We observed that a sequential training favors converging to a MF-RPN model that is nearly identical to the SF-HF-RPN model since the last learning step in fitting the MF-RPN model is nearly identical to the SF-HF-RPN model's learning of the mapping between high-fidelity parameterization inputs and outputs.

Another important aspect regarding the chosen architecture of the MF-RPN model is the uncertainty propagation across fidelity levels. Once properly trained, any uncertainty in the parameterization input is propagated to the corresponding low-fidelity parameterization output via the low-fidelity RPN (red network in figure 1). These low-fidelity predictions are directly taken as inputs for the second ensemble (blue network in figure 1). Hence, their corresponding uncertainty is naturally propagated to the corresponding high-fidelity parameterization output predictions, ensuring a continuous uncertainty propagation from the low- to high-fidelity variables.

Machine learning models usually need to be trained on normalized data. In this work, standardization (or Z-score) is used

as normalization so that inputs and outputs have the properties of a Gaussian distribution with a zero mean and unit variance. Since the MF model aggregates two datasets of different fidelity levels and therefore with different distribution supports, a choice regarding the data normalization for the MF model has to be made. On one hand, the MF model is designed to tackle the task of extrapolation beyond the high-fidelity training data. If the latter is chosen for data normalization, then the MF model would be required to make predictions for high-fidelity testing data points, while mapping them with respect to the distribution of the high-fidelity training data corresponding to the SPCAM5 historical run. On the other hand, the low-fidelity training data was built such that it provides the MF model with useful information regarding the extrapolation scenarios. If the MF-RPN model is built based on data normalization using the low-fidelity training data, then its predictions for high-fidelity testing data points will be estimated while mapping testing inputs and outputs with respect to the distribution of the low-fidelity training data corresponding to the CAM5 +8K simulation. This latter scenario of a warmer climate is closer to the high-fidelity extrapolation scenario of interest. As such, the MF-RPN model is trained on data that is normalized with a standardization based on the mean and standard deviation of the low-fidelity data corresponding to the CAM5 +8K run since the extrapolation to a warmer climate is more critical than the data accuracy. In this work, all data are normalized to unit normal distribution. Hence, MF-RPN's inputs and outputs are normalized with respect to the mean and standard deviation of the CAM5 +8K simulation dataset. This normalization applies to the high-fidelity training data: SPCAM5 historical run, and also to the low-fidelity training data: CAM5 +8K run. The chosen normalization helps the MF-RPN model account for the distribution shift between the training and testing high-fidelity data, based only on information of the computationally cheaper but valuable (for extrapolation) low-fidelity training data. Distributions of the normalized test data using statistics from CAM5 +8K and SPCAM5 historical runs confirm the physically-based motivation of using the former dataset for MF-RPN normalization as detailed above (table 1). Indeed, the normalized test data based on the CAM5 +8K statistics shows variables distributions that are closer to the unit normal one which is the ideal distribution to train the ML-based MF-RPN model on.

### RPNs' individual networks hyperparameters and training

Hyperparameters of individual neural networks forming different RPNs models did not need to be tuned from scratch, and were instead chosen based on the hyperparameter optimization over  $\sim 250$  trials conducted in Mooers et al.'s study on fully-connected neural network convection superparameterization for SPCAM5<sup>30</sup>. In particular, individual Multi-Layer Perceptrons (MLPs) forming the RPN were considered as fully connected neural networks with 7 hidden layers, each

	CAM5 +8K statistics	SPCAM5 historical run statistics
Mean of relative humidity	-0.33	1.96
Std. dev. of relative humidity	0.61	2.79
Mean of heat tendency	0.01	-0.006
Std. dev. of heat tendency	0.94	1.21

**Table 1.** Mean and standard deviation of relative humidity and heat tendency for the normalized test data using statistics from CAM5 +8K and SPCAM5 historical runs. Results are averaged across all vertical levels.

containing 512 neurons. We utilized a batch size of 2048 and ReLU activation (with a negative slope of 0.15) for all layers except for the output one, where the linear activation function was used.

The MLPs were trained for a total of 236520 stochastic gradient descent (SGD) steps using the Adam optimizer. The learning rate was initialized at  $10^{-4}$  with an exponential decay at a rate of 0.99 per 1000 steps. For data bootstrapping, each network in the RPN ensembles is trained on a randomly sampled subset with a size equal to 80% of the whole training dataset size as justified in *Yang et al.*<sup>68</sup>.

### Error metrics

In this section, we define the different error metrics that were used to evaluate the performance of the different surrogate models. We keep the formulation as generic as possible with respect to all parameterization output variables. We also keep the definition general so that it can accommodate the evaluation either on the whole test dataset (points concatenated across space and time) or on a subset of the test dataset (with concatenation along time and/or some specific space dimensions). Global error metrics will be evaluated across all test data points concatenated over space and time. For longitude-latitude structure, the error metrics are evaluated on points concatenated across time. For pressure-latitude structure, the error metrics are evaluated on points concatenated over time and longitude. Models errors are evaluated on daily averages as performed in *Mooers et al.*<sup>30</sup> in order to have a comprehensive assessment of the models performance. For the MAE metric, heat and moisture tendencies are scaled by the specific heat capacity of air at a constant pressure ( $1004.6\text{J}.\text{kg}^{-1}.\text{K}^{-1}$ ) and latent heat of vaporization at standard atmospheric conditions ( $2.26 \times 10^6\text{J}.\text{kg}^{-1}$ ), respectively<sup>30</sup>. In the next section  $y$  denotes the true target value and  $\hat{y}$  the corresponding prediction.  $\mathcal{D}$  will denote the test dataset and  $|\mathcal{D}|$  its size.

#### Mean Absolute Error (MAE):

$$\text{MAE} = \frac{1}{|\mathcal{D}|} \sum_{i \in \mathcal{D}} |y_i - \hat{y}_i|, \quad (2)$$

where  $\mathcal{D}$  denotes the test dataset,  $y_i$  the true target and  $\hat{y}_i$  the corresponding model prediction.

For global error evaluation,  $\mathcal{D}$  corresponds to the whole test dataset (points concatenated across space and time). For the longitude-latitude plots,  $\mathcal{D}$  corresponds to the test dataset concatenated across time, providing a single error metric eval-

uation for each parameterization output variable and for each point in longitude-latitude cross-section.

For pressure-latitude plots,  $\mathcal{D}$  corresponds to the test dataset concatenated across time and longitude dimension. Hence, for these plots, each pressure level corresponds to a specific parameterization output variable, and each point in latitude has a single error metric evaluation for each pressure level. For the temporal error evaluation,  $\mathcal{D}$  corresponds to the test dataset concatenated across longitude and latitude dimensions.

#### Coefficient of Determination ( $R^2$ ):

$$R^2 = 1 - \frac{\sum_{i \in \mathcal{D}} (y_i - \hat{y}_i)^2}{\sum_{i \in \mathcal{D}} (y_i - \bar{y})^2} \quad (3)$$

where  $\bar{y}$  represents the true target value averaged over the test dataset  $\mathcal{D}$ . The definition of the different choices for the test dataset  $\mathcal{D}$  is the same as detailed above for MAE.

#### Stochastic Metric (CRPS):

The Continuous Ranked Probability Score (CRPS)<sup>83,84</sup> is a generalization of the MAE for distributional predictions. CRPS penalizes over-confidence in addition to inaccuracy in ensemble predictions. A lower CRPS value corresponds to a more accurate and/or less over-confident model. For each variable, it measures the discrepancy between the ground truth target  $y$  with the cumulative distribution function (CDF)  $\hat{F}$  of the prediction via:

$$\begin{aligned} \text{CRPS}(\hat{F}, y) &= \int (\hat{F}(z) - \mathbf{1}_{\{z \geq y\}})^2 dz \\ &= \mathbb{E}[|\hat{y} - y|] - \frac{1}{2} \mathbb{E}[|\hat{y} - \hat{y}'|] \end{aligned} \quad (4)$$

where  $\hat{y}, \hat{y}' \sim \hat{F}$  are independent and identically distributed (*iid*) samples from the predicted CDF. We use the following non-parametric estimate form of the CRPS<sup>85</sup>:

$$\text{CRPS}(\hat{y}, y) = \frac{1}{n} \sum_{i=1}^n |\hat{y}_i - y| - \frac{1}{2n(n-1)} \sum_{i=1}^n \sum_{j=1}^n |\hat{y}_i - \hat{y}_j|, \quad (5)$$

where the CDF  $\hat{F}$  is estimated empirically using  $n = 32$  *iid* samples  $\hat{y}_i \sim \hat{F}$ . Equation (5) corresponds to the CRPS estimate for a singular datapoint. For a given test dataset  $\mathcal{D}$ , the corresponding CRPS is obtained as an average of individual CRPS estimates (5) over all datapoints within  $\mathcal{D}$ . The first term in (5) is the MAE between the target and samples of the

predictive distribution, while the second term is smaller for smaller predictive variances and vanishes completely for point estimates. The CRPS definition is naturally extended to the ensemble models by taking each ensemble member prediction as a sample of an implicit predictive distribution.

## References

1. Brenowitz, N. D. & Bretherton, C. S. Prognostic validation of a neural network unified physics parameterization. *Geophys. Res. Lett.* **45**, 6289–6298 (2018).
2. Rasp, S., Pritchard, M. S. & Gentine, P. Deep learning to represent subgrid processes in climate models. *Proc. Natl. Acad. Sci.* **115**, 9684–9689 (2018).
3. Gentine, P., Pritchard, M., Rasp, S., Reinaudi, G. & Yacalis, G. Could machine learning break the convection parameterization deadlock? *Geophys. Res. Lett.* **45**, 5742–5751 (2018).
4. O’Gorman, P. A. & Dwyer, J. G. Using machine learning to parameterize moist convection: Potential for modeling of climate, climate change, and extreme events. *J. Adv. Model. Earth Syst.* **10**, 2548–2563 (2018).
5. Bolton, T. & Zanna, L. Applications of deep learning to ocean data inference and subgrid parameterization. *J. Adv. Model. Earth Syst.* **11**, 376–399 (2019).
6. Ross, A., Li, Z., Perezhigin, P., Fernandez-Granda, C. & Zanna, L. Benchmarking of machine learning ocean subgrid parameterizations in an idealized model. *J. Adv. Model. Earth Syst.* **15**, e2022MS003258 (2023). E2022MS003258 2022MS003258.
7. Partee, S. *et al.* Using machine learning at scale in numerical simulations with smartsim: An application to ocean climate modeling. *J. Comput. Sci.* **62**, 101707 (2022).
8. Couvreux, F. *et al.* Process-based climate model development harnessing machine learning: I. a calibration tool for parameterization improvement. *J. Adv. Model. Earth Syst.* **13**, e2020MS002217 (2021). E2020MS002217 2020MS002217.
9. Kashinath, K. *et al.* Physics-informed machine learning: case studies for weather and climate modelling. *Philos. Transactions Royal Soc. A: Math. Phys. Eng. Sci.* **379**, 20200093 (2021).
10. Zanna, L. & Bolton, T. *Deep Learning of Unresolved Turbulent Ocean Processes in Climate Models*, chap. 20, 298–306 (John Wiley & Sons, Ltd, 2021).
11. Sonnewald, M. *et al.* Bridging observations, theory and numerical simulation of the ocean using machine learning. *Environ. Res. Lett.* **16**, 073008 (2021).
12. Lorenz, E. Predictability: a problem partly solved. In *Seminar on Predictability, 4-8 September 1995*, vol. 1, 1–18. ECMWF (ECMWF, Shinfield Park, Reading, 1995).
13. Palmer, T. N. A nonlinear dynamical perspective on model error: A proposal for non-local stochastic-dynamic parametrization in weather and climate prediction models. *Q. J. Royal Meteorol. Soc.* **127**, 279–304 (2001).
14. Tribbia, J. J. & Baumhefner, D. P. Scale interactions and atmospheric predictability: An updated perspective. *Mon. Weather. Rev.* **132**, 703 – 713 (2004).
15. Karimi, A. & Paul, M. Extensive chaos in the lorenz-96 model. *Chaos (Woodbury, N.Y.)* **20**, 043105 (2010).
16. Palmer, T. N. Towards the probabilistic earth-system simulator: a vision for the future of climate and weather prediction. *Q. J. Royal Meteorol. Soc.* **138**, 841–861 (2012).
17. Wang, Y., Zhang, G. J. & Craig, G. C. Stochastic convective parameterization improving the simulation of tropical precipitation variability in the ncar cam5. *Geophys. Res. Lett.* **43**, 6612–6619 (2016).
18. Davini, P. *et al.* Climate sphinx: evaluating the impact of resolution and stochastic physics parameterisations in the ec-earth global climate model. *Geosci. Model. Dev.* **10**, 1383–1402 (2017).
19. Christensen, H. M., Berner, J., Coleman, D. R. B. & Palmer, T. N. Stochastic parameterization and el niño-southern oscillation. *J. Clim.* **30**, 17 – 38 (2017).
20. Strømmen, K., Christensen, H., Berner, J. & Palmer, T. The impact of stochastic parametrisations on the representation of the asian summer monsoon. *Clim. Dyn.* **50** (2018).
21. Berner, J., Jung, T. & Palmer, T. N. Systematic model error: The impact of increased horizontal resolution versus improved stochastic and deterministic parameterizations. *J. Clim.* **25**, 4946 – 4962 (2012).
22. Seiffert, R. & von Storch, J.-S. A stochastic analysis of the impact of small-scale fluctuations on the tropospheric temperature response to co2 doubling. *J. Clim.* **23**, 2307 – 2319 (2010).
23. Ajayamohan, R. S., Khouider, B. & Majda, A. J. Realistic initiation and dynamics of the madden-julian oscillation in a coarse resolution aquaplanet gcm. *Geophys. Res. Lett.* **40**, 6252–6257 (2013).
24. Dawson, A. & Palmer, T. N. Simulating weather regimes: impact of model resolution and stochastic parameterization. *Clim. Dyn.* **44**, 2177–2193 (2015).
25. Berner, J. *et al.* Stochastic parameterization: Toward a new view of weather and climate models. *Bull. Am. Meteorol. Soc.* **98**, 565 – 588 (2017).
26. Brenowitz, N. D., Beucler, T., Pritchard, M. & Bretherton, C. S. Interpreting and stabilizing machine-learning parameterizations of convection. *J. Atmos. Sci.* **77**, 4357–4375 (2020).

- 27.** Han, Y., Zhang, G. J., Huang, X. & Wang, Y. A moist physics parameterization based on deep learning. *J. Adv. Model. Earth Syst.* **12**, e2020MS002076 (2020).
- 28.** Ott, J. *et al.* A fortran-keras deep learning bridge for scientific computing. *arXiv preprint arXiv:2004.10652* (2020).
- 29.** Iglesias-Suarez, F. *et al.* Causally-informed deep learning to improve climate models and projections. *arXiv preprint arXiv:2304.12952* (2023).
- 30.** Mooers, G. *et al.* Assessing the potential of deep learning for emulating cloud superparameterization in climate models with real-geography boundary conditions. *J. Adv. Model. Earth Syst.* **13**, e2020MS002385 (2021). E2020MS002385 2020MS002385.
- 31.** Wang, X., Han, Y., Xue, W., Yang, G. & Zhang, G. J. Stable climate simulations using a realistic general circulation model with neural network parameterizations for atmospheric moist physics and radiation processes. *Geosci. Model. Dev.* **15**, 3923–3940 (2022).
- 32.** Frezat, H., Le Sommer, J., Fablet, R., Balarac, G. & Lguensat, R. A posteriori learning for quasi-geostrophic turbulence parametrization. *J. Adv. Model. Earth Syst.* **14**, e2022MS003124 (2022). E2022MS003124 2022MS003124.
- 33.** Krasnopolsky, V. M., Fox-Rabinovitz, M. S. & Chalikov, D. V. New approach to calculation of atmospheric model physics: Accurate and fast neural network emulation of longwave radiation in a climate model. *Mon. Weather Rev.* **133**, 1370 – 1383 (2005).
- 34.** Schneider, T., Lan, S., Stuart, A. & Teixeira, J. a. Earth system modeling 2.0: A blueprint for models that learn from observations and targeted high-resolution simulations. *Geophys. Res. Lett.* **44**, 12,396–12,417 (2017).
- 35.** Gettelman, A. *et al.* Machine learning the warm rain process. *J. Adv. Model. Earth Syst.* **13**, e2020MS002268 (2021). E2020MS002268 2020MS002268.
- 36.** Bretherton, C. S. *et al.* Correcting coarse-grid weather and climate models by machine learning from global storm-resolving simulations. *J. Adv. Model. Earth Syst.* **14**, e2021MS002794 (2022).
- 37.** Clark, S. K. *et al.* Correcting a 200 km resolution climate model in multiple climates by machine learning from 25 km resolution simulations. *J. Adv. Model. Earth Syst.* **14**, e2022MS003219 (2022).
- 38.** Kwa, A. *et al.* Machine-learned climate model corrections from a global storm-resolving model: Performance across the annual cycle. *J. Adv. Model. Earth Syst.* **15**, e2022MS003400 (2023).
- 39.** Sanford, C. H. *et al.* Improving the reliability of ml-corrected climate models with novelty detection. *Austraea Prepr.* (2023).
- 40.** Fernandez-Godino, M. G. Review of multi-fidelity models. *arXiv preprint arXiv:1609.07196v4* (2023).
- 41.** Perdikaris, P., Raissi, M., Damianou, A., Lawrence, N. D. & Karniadakis, G. E. Nonlinear information fusion algorithms for data-efficient multi-fidelity modelling. *Proc. Royal Soc. A: Math. Phys. Eng. Sci.* **473**, 20160751 (2017).
- 42.** Meng, X., Babaee, H. & Karniadakis, G. E. Multi-fidelity bayesian neural networks: Algorithms and applications. *J. Comput. Phys.* **438**, 110361 (2021).
- 43.** Chen, C., Zuo, Y., Ye, W., Li, X. & Ong, S. P. Learning properties of ordered and disordered materials from multi-fidelity data. *Nat. Comput. Sci.* **1**, 46–53 (2021).
- 44.** Liu, D. & Wang, Y. Multi-Fidelity Physics-Constrained Neural Network and Its Application in Materials Modeling. *J. Mech. Des.* **141**, 121403 (2019).
- 45.** Zhang, X., Xie, F., Ji, T., Zhu, Z. & Zheng, Y. Multi-fidelity deep neural network surrogate model for aerodynamic shape optimization. *Comput. Methods Appl. Mech. Eng.* **373**, 113485 (2021).
- 46.** Wu, D., Chinazzi, M., Vespignani, A., Ma, Y.-A. & Yu, R. Multi-fidelity hierarchical neural processes. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, KDD '22, 2029–2038 (Association for Computing Machinery, New York, NY, USA, 2022).
- 47.** Peherstorfer, B., Willcox, K. & Gunzburger, M. Survey of multifidelity methods in uncertainty propagation, inference, and optimization. *SIAM Rev.* **60**, 550–591 (2018).
- 48.** Fernandez-Godino, M. G., Park, C., Kim, N.-H. & Haftka, R. T. Review of multi-fidelity models. *arXiv preprint arXiv:1609.07196* (2017).
- 49.** Neal, R. M. *et al.* Mcmc using hamiltonian dynamics. *Handb. markov chain monte carlo* **2**, 2 (2011).
- 50.** Hinton, G. E. & van Camp, D. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, COLT '93, 5–13 (Association for Computing Machinery, New York, NY, USA, 1993).
- 51.** Blundell, C., Cornebise, J., Kavukcuoglu, K. & Wierstra, D. Weight uncertainty in neural networks. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, 1613–1622 (JMLR.org, 2015).
- 52.** Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, 6405–6416 (Curran Associates Inc., Red Hook, NY, USA, 2017).

53. Fort, S., Hu, H. & Lakshminarayanan, B. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757v2* (2019).
54. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: A simple way to prevent neural networks from overfitting. *J. Mach. Learn. Res.* **15**, 1929–1958 (2014).
55. Gal, Y. & Ghahramani, Z. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In Balcan, M. F. & Weinberger, K. Q. (eds.) *Proceedings of The 33rd International Conference on Machine Learning*, vol. 48 of *Proceedings of Machine Learning Research*, 1050–1059 (PMLR, New York, New York, USA, 2016).
56. Osband, I. *et al.* Evaluating predictive distributions: Does bayesian deep learning work? (2022).
57. Riquelme, C., Tucker, G. & Snoek, J. Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling. In *International Conference on Learning Representations* (2018).
58. Osband, I., Aslanides, J. & Cassirer, A. Randomized prior functions for deep reinforcement learning. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS’18, 8626–8638 (Curran Associates Inc., Red Hook, NY, USA, 2018).
59. Yang, Y., Bhouri, M. A. & Perdikaris, P. Bayesian differential programming for robust systems identification under uncertainty. *Proc. Royal Soc. A: Math. Phys. Eng. Sci.* **476**, 20200290 (2020).
60. Bhouri, M. A., Joly, M., Yu, R., Sarkar, S. & Perdikaris, P. Scalable bayesian optimization with high-dimensional outputs using randomized prior networks. *arXiv preprint arXiv:2302.07260v4* (2023).
61. Ciosek, K., Fortuin, V., Tomioka, R., Hofmann, K. & Turner, R. Conservative uncertainty estimation by fitting prior networks. In *International Conference on Learning Representations* (2020).
62. Behrens, G. *et al.* Non-linear dimensionality reduction with a variational encoder decoder to understand convective processes in climate models. *J. Adv. Model. Earth Syst.* **14**, e2022MS003130 (2022).
63. Yu, S. *et al.* Climsim: An open large-scale dataset for training high-resolution physics emulators in hybrid multi-scale climate simulators. *arXiv preprint arXiv:2306.08754* (2023).
64. Grabowski, W. W. Coupling cloud processes with the large-scale dynamics using the cloud-resolving convection parameterization (crcp). *J. Atmospheric Sci.* **58**, 978 – 997 (2001).
65. Randall, D., Khairoutdinov, M., Arakawa, A. & Grabowski, W. Breaking the cloud parameterization deadlock. *Bull. Am. Meteorol. Soc.* **84**, 1547 – 1564, DOI: <https://doi.org/10.1175/BAMS-84-11-1547> (2003).
66. Iacono, M. J. *et al.* Radiative forcing by long-lived greenhouse gases: Calculations with the aer radiative transfer models. *J. Geophys. Res. Atmospheres* **113** (2008).
67. Mlawer, E. J., Taubman, S. J., Brown, P. D., Iacono, M. J. & Clough, S. A. Radiative transfer for inhomogeneous atmospheres: Rrtm, a validated correlated-k model for the longwave. *J. Geophys. Res. Atmospheres* **102**, 16663–16682 (1997).
68. Yang, Y., Kissas, G. & Perdikaris, P. Scalable uncertainty quantification for deep operator networks using randomized priors. *Comput. Methods Appl. Mech. Eng.* **399**, 115399 (2022).
69. Glorot, X. & Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In Teh, Y. W. & Titterington, M. (eds.) *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, vol. 9 of *Proceedings of Machine Learning Research*, 249–256 (PMLR, Chia Laguna Resort, Sardinia, Italy, 2010).
70. Kwa, A. *et al.* Machine-learned climate model corrections from a global storm-resolving model. *arXiv preprint arXiv:2211.11820* (2022).
71. Watt-Meyer, O. *et al.* Correcting weather and climate models by machine learning nudged historical simulations. *Geophys. Res. Lett.* **48**, e2021GL092555 (2021).
72. Oueslati, B. & Bellon, G. The double ITCZ bias in CMIP5 models: interaction between SST, large-scale circulation and precipitation. *Clim. Dyn.* **44**, 585–607 (2015).
73. Bador, M. *et al.* Impact of higher spatial atmospheric resolution on precipitation extremes over land in global climate models. *J. Geophys. Res. Atmospheres* **125**, e2019JD032184 (2020). E2019JD032184 2019JD032184.
74. Hung, M.-P. *et al.* Mjo and convectively coupled equatorial waves simulated by cmip5 climate models. *J. Clim.* **26**, 6185 – 6214 (2013).
75. Beucler, T. *et al.* Enforcing analytic constraints in neural networks emulating physical systems. *Phys. Rev. Lett.* **126**, 098302 (2021).
76. Reed, C. J. *et al.* Scale-mae: A scale-aware masked autoencoder for multiscale geospatial representation learning. *arXiv preprint arXiv:2212.14532* (2023).
77. Bhouri, M. A. & Gentine, P. Memory-based parameterization with differentiable solver: Application to lorenz ’96. *Chaos: An Interdiscip. J. Nonlinear Sci.* **33**, 073116 (2023).
78. Zhang, Y. & Chen, H. Comparing cam5 and superparameterized cam5 simulations of summer precipitation

- characteristics over continental east asia: Mean state, frequency-intensity relationship, diurnal cycle, and influencing factors. *J. Clim.* **29**, 1067–1089 (2016).
79. Morrison, H. & Gettelman, A. A new two-moment bulk stratiform cloud microphysics scheme in the community atmosphere model, version 3 (cam3). part i: Description and numerical tests. *J. Clim.* **21**, 3642 – 3659 (2008).
  80. Park, S. & Bretherton, C. S. The university of washington shallow convection and moist turbulence schemes and their impact on climate simulations with the community atmosphere model. *J. Clim.* **22**, 3449 – 3469 (2009).
  81. Bretherton, C. S. & Park, S. A new moist turbulence parameterization in the community atmosphere model. *J. Clim.* **22**, 3422 – 3448 (2009).
  82. Hurrell, J. W., Hack, J. J., Shea, D., Caron, J. M. & Rosinski, J. A new sea surface temperature and sea ice boundary dataset for the community atmosphere model. *J. Clim.* **21**, 5145 – 5153 (2008).
  83. Matheson, J. E. & Winkler, R. L. Scoring rules for continuous probability distributions. *Manag. Sci.* **22**, 1087–1096 (1976).
  84. Gneiting, T. & Raftery, A. E. Strictly proper scoring rules, prediction, and estimation. *J. Am. Stat. Assoc.* **102**, 359–378 (2007).
  85. Ferro, C. A. T. Fair scores for ensemble forecasts. *Q. J. R. Meteorol. Soc.* **140**, 1917–1923 (2014).

## Acknowledgements

The authors would like to acknowledge funding from the National Science Foundation LEAP Science and Technology center Award # 2019625, USMILE European Research Council (ERC) Synergy grant, National Science Foundation funding from an AGS-PRF Fellowship Award (AGS2218197) and Department of Energy.

## Author contributions statement

M.A.B. and P.G. conceived the experiment(s), L.P. generated the training data with guidance by M.P., M.A.B. conducted the experiment(s), M.A.B. and P.G. analyzed the results, M.A.B., P.G. and M.P. provided funding, M.A.B., P.G., L.P. and M.P. wrote the manuscript.

## Additional information

Following paper publication, all code and data accompanying this manuscript will be made publicly available at [https://github.com/bhouria0412/MF\\_RPN\\_SPCAM\\_cv\\_param](https://github.com/bhouria0412/MF_RPN_SPCAM_cv_param).

Authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

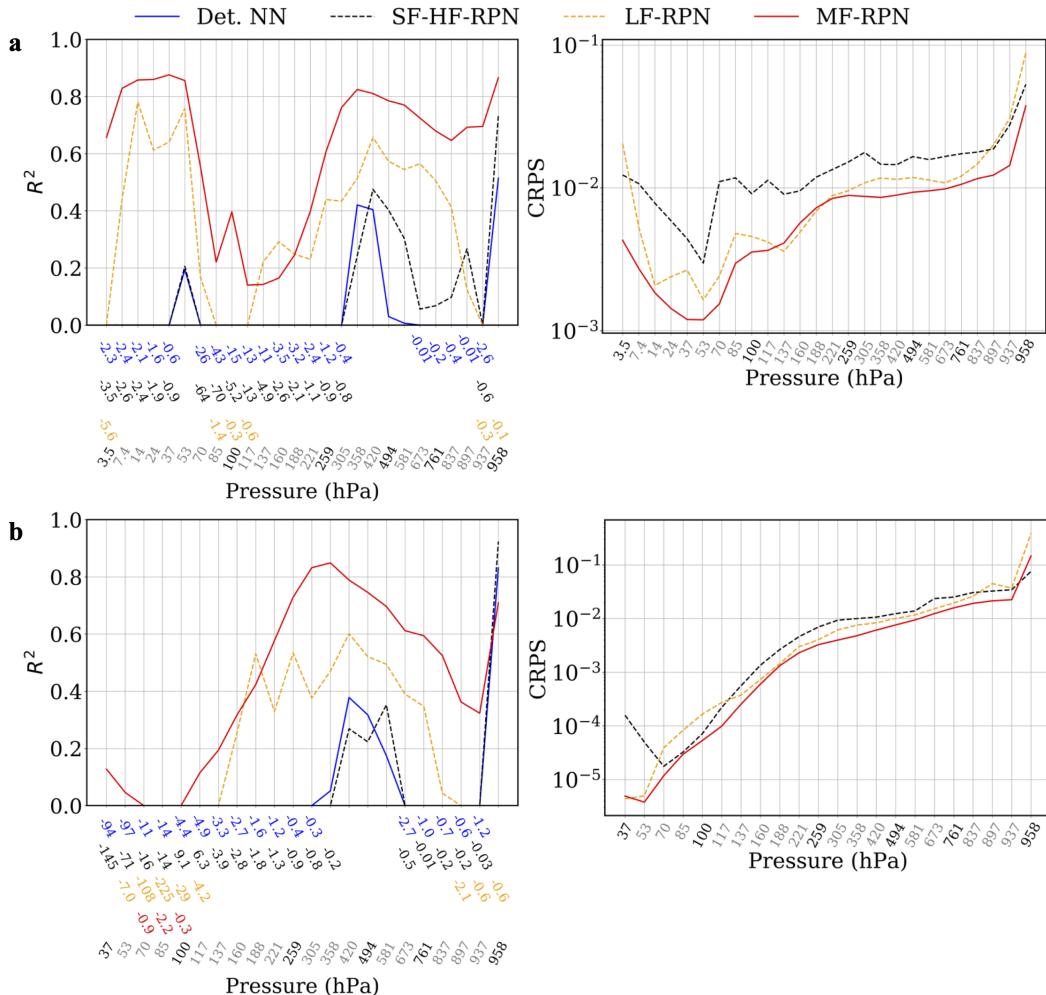
## Supplementary Information

### Global errors

For the heat tendency and the vertical levels where the deterministic NN shows significantly negative  $R^2$  values, the SF-HF-RPN  $R^2$  values are also negative but very similar to those of the deterministic NN (between 3.5 and 305 hPa, figure 6.a). However, for the vertical levels where the deterministic NN shows nearly zero to slightly negative  $R^2$  values (from 494 to 897 hPa), the SF-HF-RPN does significantly improve upon the deterministic NN, showing the ability of the RPN surrogate model to resolve part of the convection stochasticity.

For the moisture tendency and as mentioned in the Results section, the negative  $R^2$  values for the deterministic NN. and SF-HF-RPN show that the latter resolves better the moisture convection stochasticity for vertical levels below 494 hPa, while it does not improve upon the deterministic NN. for higher levels (figure 6.b).

The CRPS curves follow a similar trend as the MAE ones (see Results), showing that all RPN-based models have similar confidence in terms of variance prediction (figure 6). This behavior could be expected since all these models are based on the same numerical ensemble technique of RPN. Nonetheless, the uncertainty quantification returned by these different RPN-based models are quite different and do not vary similarly as a function of the actual error (see Uncertainty quantification).



**Figure 6.  $R^2$  and CRPS for different models across all test data points concatenated over space and time.** Y-axes of  $R^2$  plots are limited to  $[0, 1]$  for clarity purposes. Negative  $R^2$  values are indicated below the plot with a  $60^\circ$  clockwise rotation and the color corresponding to the one used to plot the specific model's results. **a**, Heat tendency results. **b**, Moisture tendency results

### Longitude-latitude errors structure

Figure 7 shows the longitude-latitude structure of the  $R^2$  and MAE metrics for moisture tendency at vertical levels 259, 494 and 761 hPa for all surrogate models and evaluated on the test dataset. For the  $R^2$  metric, the moisture tendency results for all pressure levels are quite similar with the SF-HF-RPN model improving upon the deterministic NN model in the temperate zone, and the LF-RPN model improving further upon the SF-HF-RPN model within the temperate zone and even within the tropics and polar regions.

For all vertical levels considered, the MF-RPN model improves even further compared to all other models across all regions, while still showing a few negative  $R^2$  values for this extrapolation task like all other models. For the pressure level 494 hPa, the negative  $R^2$  regions for the MF-RPN model include part of the South-East Pacific region, of the south Atlantic ocean and of the African Sahara. However, all these regions are significantly smaller for the pressure level 259 hPa compared to the regions observed for 494 hPa (white regions in MF-RPN plots in figures 7.a and 7.b). Similarly the negative  $R^2$  regions for the MF-RPN model within the South-East Pacific region and south Atlantic ocean are clearly smaller at the pressure level 761 hPa compared to level 494 hPa (white regions in MF-RPN plots in figures 7.b and 7.c). Hence, the MF-RPN model is capable of better extrapolating at relatively low- and high-altitude levels within the troposphere (259 and 761 hPa) compared to mid-altitude levels (around 494 hPa).

Since the highest error is observed within the tropical regions, the MAE longitude-latitude plots (figures 7.d - 7.f) give a clear representation of the performance improvement within these regions across different models. For all pressure levels considered, the LF-RPN systematically improves the moisture convection parameterization within the tropics compared to the deterministic NN. and SF-HF-RPN models, while the MF-RPN improves even further upon the LF-RPN model. Comparing the deterministic NN. and SF-HF-RPN results, the latter is capable of better resolving the moisture convection stochasticity within the tropics for low vertical levels (761 hPa as seen in figure 7.f), while it fails to do so for higher levels (259 and 494 hPa as seen in figures 7.d and 7.e).

Figure 8 shows the longitude-latitude structure of the  $R^2$  and MAE metrics for heat tendency at vertical levels 259, 494 and 761 hPa for all surrogate models and evaluated on the test dataset. For the  $R^2$  variation, the heat tendency results for all pressure levels are quite similar to the results obtained for the moisture tendency, with the SF-HF-RPN model improving upon the deterministic NN. one in the temperate zone, and the LF-RPN model improving further upon the SF-HF-RPN model within the temperate zone and even within the tropics and polar regions.

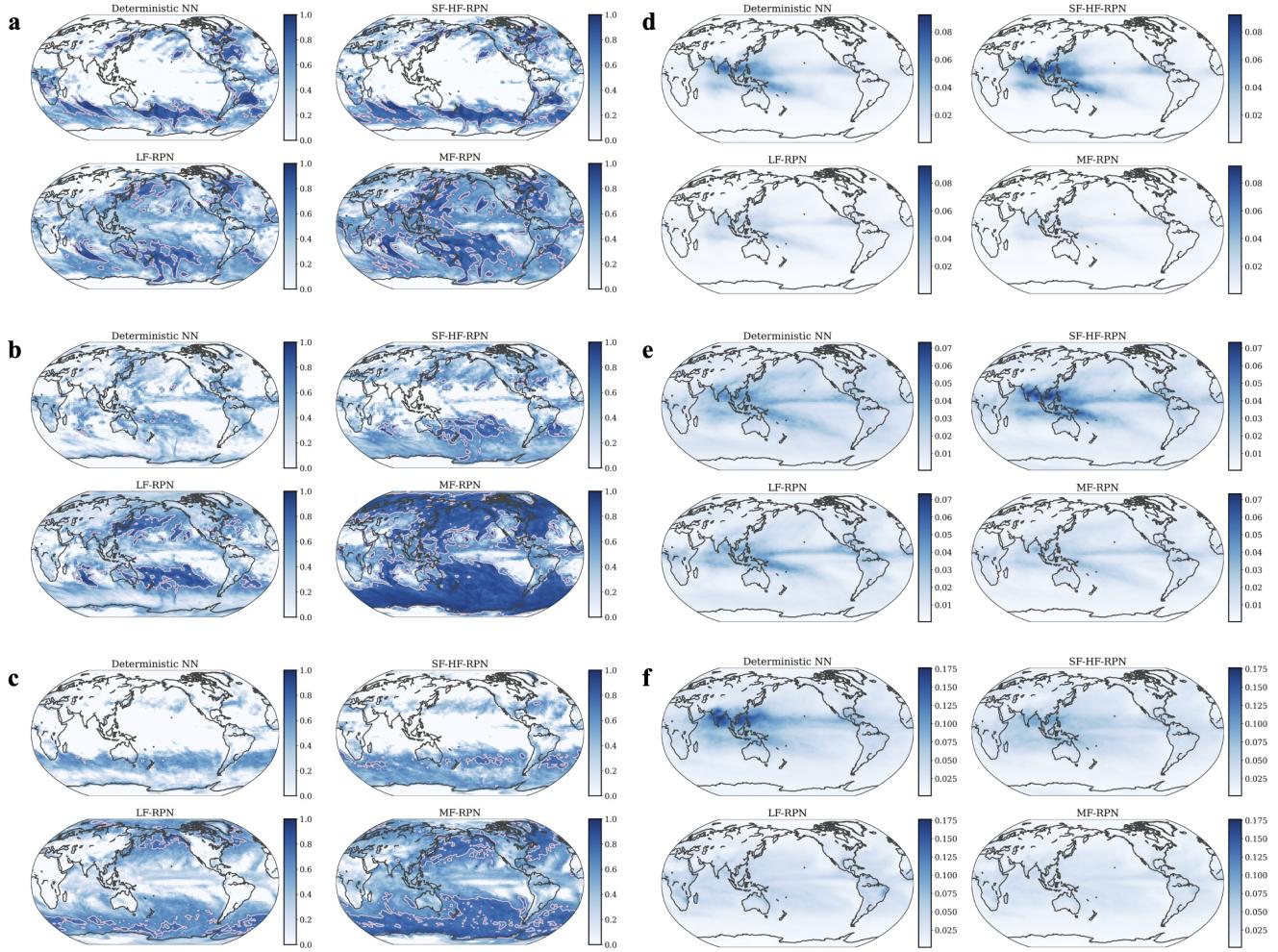
For all vertical levels considered, the MF-RPN model improves even further compared to all other models across all regions, while still showing some negative  $R^2$  values for this extrapolation task like all other models. For the pressure level 494 hPa, the negative  $R^2$  regions for the MF-RPN model include part of the South-East Pacific region, of the south Atlantic ocean and of the African Sahara as observed for the moisture tendency. However, all these regions are significantly smaller and nearly non-existent for the pressure levels 259 and 761 hPa compared to the regions observed for 494 hPa (white regions in MF-RPN plots in figures 8.a-8.c). Hence, the MF-RPN model is capable of better extrapolating at relatively low and high altitude levels within the troposphere (259 and 761 hPa) compared to mid altitude levels (around 494 hPa) as observed for the moisture tendency.

It is worth noting that at vertical level 494 hPa, the LF-RPN model shows better results compared to the MF-RPN model in the South-East Pacific region, south Atlantic ocean and the African Sahara (figure 8.b). In these regions, the LF-RPN model returns overall positive  $R^2$  values, while the MF-RPN fails to do so. This result suggests that the LF-RPN model is better at approximating shallow convection, while the MF-RPN model does a better job at learning deep convection. The difference in behavior can be attributed to the different datasets on which these two models have been trained since the LF-RPN is only trained on the 1-year CAM5 +8K simulation, while the MF-RPN model has to aggregate both the latter dataset with the 3-month SPCAM5 historical run. Nonetheless, at the vertical level of 494 hPa, the MF-RPN clearly outperforms the LF-RPN model in all other tropical regions except the three ones listed above showing again the overall better generalization and extrapolation achieved by the MF model.

The MAE longitude-latitude plots (figures 8.d - 8.f) clearly show the improvement obtained for the heat convection parameterization with the MF-RPN compared to the LF-RPN for all pressure levels considered. The LF-RPN model also improves upon the deterministic NN. and SF-HF-RPN models. However, unlike the moisture tendency results, the SF-HF-RPN model does not show higher errors compared to the deterministic NN. model at pressure level 259 hPa (figure 8.d), and does improve the heat convection parameterization at pressure levels 494 and 761 hPa (figures 8.e and 8.f). Hence, the SF-HF-RPN is capable of better resolving the convection stochasticity across different vertical levels for the heat tendency compared to the moisture one.

### Pressure-latitude errors structure

Based on the data distribution of moisture tendency for the lowest vertical level at 958 hPa (figure 9.b), it is quite obvious that for this particular level, the low-fidelity training data of CAM5 +8K simulation is not informative on the extrapolation scenario



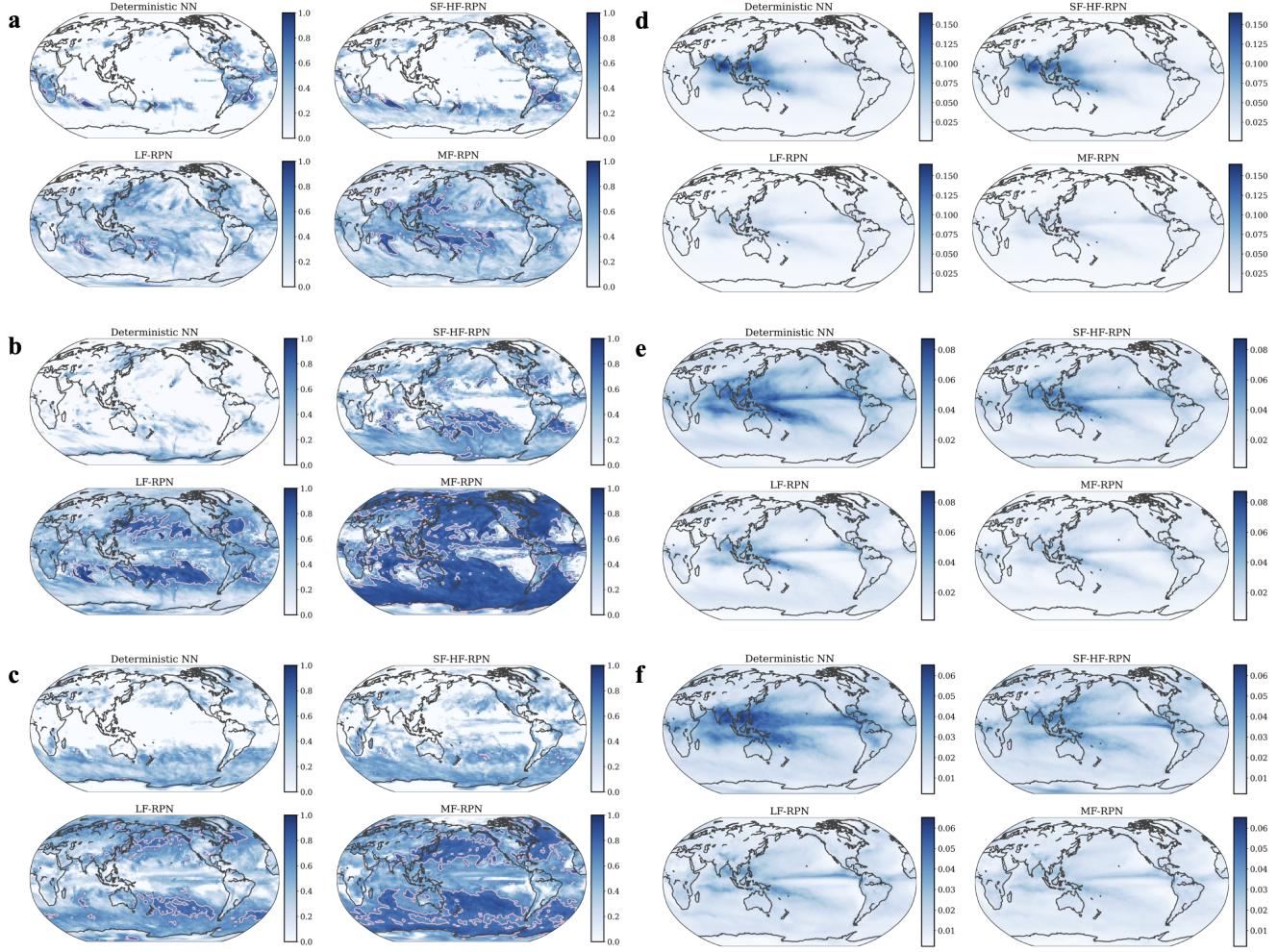
**Figure 7. Longitude-latitude variation of MAE and  $R^2$  for moisture tendency at different vertical levels.**  $R^2$  is evaluated on the test dataset and negative values are lumped to 0 for clarity purposes. **a**,  $R^2$  at  $P = 259$  hPa. **b**,  $R^2$  at  $P = 494$  hPa. **c**,  $R^2$  at  $P = 761$  hPa. **d**, MAE at  $P = 259$  hPa. **e**, MAE at  $P = 494$  hPa. **f**, MAE at  $P = 761$  hPa. It is an encouraging sign of successful stochastic learning that RPN predictions are most uncertain in the same tropical locations where deterministic errors are highest.

of interest when compared to data distributions of HF training and testing data. Similarly, the CAM5 +4K is not informative neither as the corresponding data spans a similar interval to the CAM5 +8K dataset. For other vertical levels, the CAM5 +8K simulation dataset does extrapolate beyond the HF training data (see Methods). These datasets distributions explain the overall lower performance observed for LF-RPN and MF-RPN models when it comes to infer the moisture tendency at the lowest vertical level 958 hPa (figure 9.a).

### Temporal errors structure

Figure 10 shows the temporal variations of testing MAE and  $R^2$  for heat tendency and different surrogate models at different vertical levels. In coherence with the global errors and spatial structures of different error metrics, the MF-RPN is the best performing model across all testing period and different vertical levels, followed by the LF-RPN model. The errors temporal variations show a better performance overall for the SF-HF-RPN model compared to the deterministic NN, mostly for the vertical level 494 hPa, while the improvement for the two other vertical levels is mainly limited to the second half of the testing period of a year.

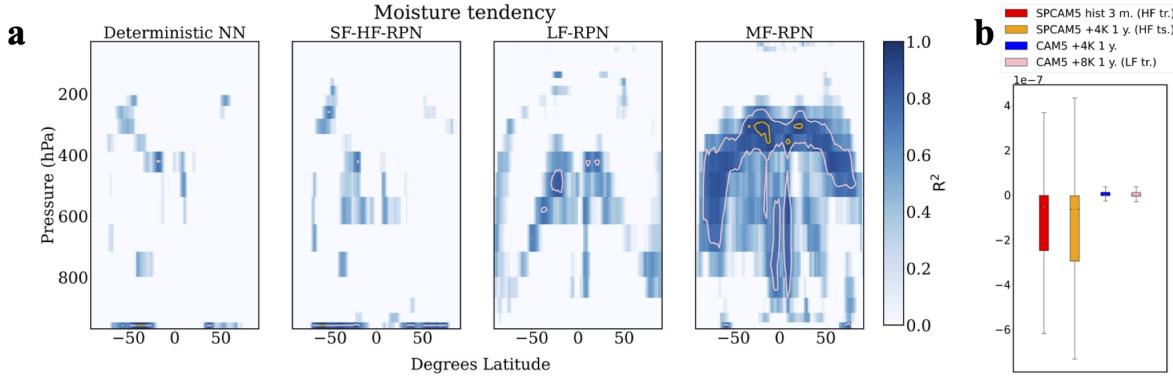
The seasonality effect on the models performance is not quite trivial. We remind that the SF-HF-RPN and deterministic NN, are trained on an SPCAM5 historical run simulation from February 1st 2003 to April 31st 2003, while the LF-RPN is trained (jointly to the MF-RPN training) on a CAM5 +8K simulation from February 1st 2003 to January 31st 2004. The MF-RPN model is trained on both datasets. The test dataset corresponds to an SPCAM5 +4K run simulation from February 1st 2003 to January 31st 2004. However, the temporal variations of the SF-HF-RPN and deterministic NN, testing errors show poor



**Figure 8. Longitude-latitude variation of MAE and  $R^2$  for heat tendency at different vertical levels.**  $R^2$  is evaluated on the test dataset and negative values are lumped to 0 for clarity purposes. **a**,  $R^2$  at  $P = 259$  hPa. **b**,  $R^2$  at  $P = 494$  hPa. **c**,  $R^2$  at  $P = 761$  hPa. **d**, MAE at  $P = 259$  hPa. **e**, MAE at  $P = 494$  hPa. **f**, MAE at  $P = 761$  hPa.

performance even in the period between February 1st 2003 and April 31st 2003 (figures 10.a and 10.d). Hence, the SF-HF-RPN and deterministic NN.'s error temporal variations for the vertical level 259 hPa are mostly governed by extrapolating to a different climate scenario rather than extrapolating to a different seasonality. On the other hand, their error temporal variations for the vertical level 494 hPa do show a significantly lower performance when extrapolating beyond the training period (figures 10.b and 10.e). Therefore, the testing error for the vertical level 494 hPa is affected not only by the extrapolation to a different climate but also to a different seasonality. The SF-HF-RPN and deterministic NN.'s error temporal variations for the vertical level 761 hPa show lower performance around the period between June and August, but the values observed for the second half of the testing year are similar to those obtained for the period between February and May (figures 10.c and 10.f). Hence, the temporal extrapolation for the vertical level 761 hPa seems to affect the error differently based on the unseen season among the training dataset. Note that since the testing dataset corresponds to a climate simulation that was not considered in any of the training datasets, it is always challenging to disentangle the different extrapolations contribution on the models errors. Given the heterogeneous temporal variations of the SF-HF-RPN and deterministic NN. errors for different seasons across different vertical levels, it is not straightforward to draw conclusions on the seasonality effect on these models performance.

Figure 11 shows the temporal variations of testing MAE and  $R^2$  for moisture tendency and different surrogate models at different vertical levels. In coherence with the global errors and spatial structures of different error metrics, the MF-RPN is the best performing model across all testing period and different vertical levels, followed by the LF-RPN model. The errors temporal variations show a better performance for the SF-HF-RPN model compared to the deterministic NN. only for the vertical level 761 hPa, while the deterministic NN. outperforms the SF-HF-RPN model for vertical levels 259 and 494 hPa. These results are well coherent with the global errors and the longitude-latitude structures of  $R^2$ , confirming that the



**Figure 9. Pressure-latitudinal variation of coefficient of determination  $R^2$  for moisture tendency and different surrogate models.**  $R^2$  is evaluated on the test dataset and negative values are lumped to 0 for clarity purposes. **a**, Moisture tendency results. **b**, Data distribution of moisture tendency for different datasets at lowest vertical level at 958 hPa.

SF-HF-RPN is more capable of resolving the heat convection stochasticity and is only able of better resolving the moisture convection stochasticity for vertical levels below the 494hPa one, while it struggles to do so for higher levels.

As observed for the heat convection parameterization, the seasonality effect on the models performance is not quite trivial. Indeed, for the vertical level 259 hPa the temporal variations of the SF-HF-RPN and deterministic NN testing errors show poor performance even in the period between February 1st 2003 and April 31st 2003 (figures 11.a and 11.d). On the other hand, their error temporal variations for the vertical levels 494 and 761 hPa do show a significantly lower performance when extrapolating beyond the training period (figures 11.b, 11.c, 11.e and 11.f). These results along with those obtained for the heat convection parameterization confirm that it is not straightforward to draw conclusions on the seasonality effect on these models performance.

### Uncertainty quantification

In addition to the deterministic and statistical error metrics, we compare the performance of different Bayesian surrogate models with respect to the returned uncertainty quantification. For a given input  $x$ , each of the RPN-based models returns an ensemble of predictions from which we infer the mean  $\hat{y}(x)$  for each of the output variables. The mean is used to estimate the different error metrics discussed above. We can also estimate the corresponding standard deviation  $\sigma_M(x)$  which serves as an uncertainty quantification. We use the subscript  $M$  to emphasize that we consider the model's inherent uncertainty that is estimated for each individual input point. Standard deterministic machine learning-based parameterizations do not provide the inherent uncertainty and can only return uncertainties by averaging over several input points (e.g. different input points across space and/or time which gives uncertainties that are "contaminated" with spatial and/or temporal physical variations).

Figure 12 shows the density plots of the uncertainty  $\sigma_M$  as a function of the prediction error evaluated on the test dataset for different RPN-based models. It includes the results for heat and moisture tendencies at vertical levels 259, 494 and 761 hPa. A perfect uncertainty quantification would strictly increase with the model's error. For the heat tendency, the MF-RPN's uncertainty clearly displays a more increasing behavior with the error compared to the SF-HF-RPN and LF-RPN models for all vertical levels since the red and yellow regions stretch over larger areas indicating a wider increase of the uncertainty as the error grows (figures 12.a - 12.c). For instance, for the vertical level  $P = 259$  hPa, the LF-RPN and SF-HF-RPN density plots clearly show narrow red regions where the returned uncertainty does not vary much compared to the error, unlike the MF-RPN's estimated uncertainty (figure 12.a).

For the moisture tendency, the MF-RPN's uncertainty still displays a more increasing behavior with the error compared to the SF-HF-RPN and LF-RPN models for all vertical levels since the yellow regions stretch over larger areas indicating a wider increase of the uncertainty as the error grows (figures 12.d - 12.f). However, it is worth noting that the SF-HF-RPN's uncertainty density plot shows wider red regions for vertical levels 494 and 761 hPa compared to the MF-RPN (figures 12.e and 12.f). Hence, the SF-HF-RPN's uncertainty displays a more increasing behavior with the error for low error values. Nonetheless, the density plots yellow regions are larger for the MF-RPN model compared to the SF-HF-RPN, highlighting that the MF-RPN's uncertainty is more accurate for moderate to high error values.

Figures 13 shows the longitude-latitude structures of the MF-RPN's uncertainty  $\sigma_M$  and the MAE metric evaluated on the testing dataset. It includes the results for heat and moisture tendencies at vertical levels 259, 494 and 761 hPa. Figures 14 and 15 provide the same results for the LF-RPN and SF-HF-RPN models respectively. Unlike the density plots and as conducted for the MAE estimates to obtain the longitude-latitude structures, the model's uncertainty  $\sigma_M$  is only averaged over time here in

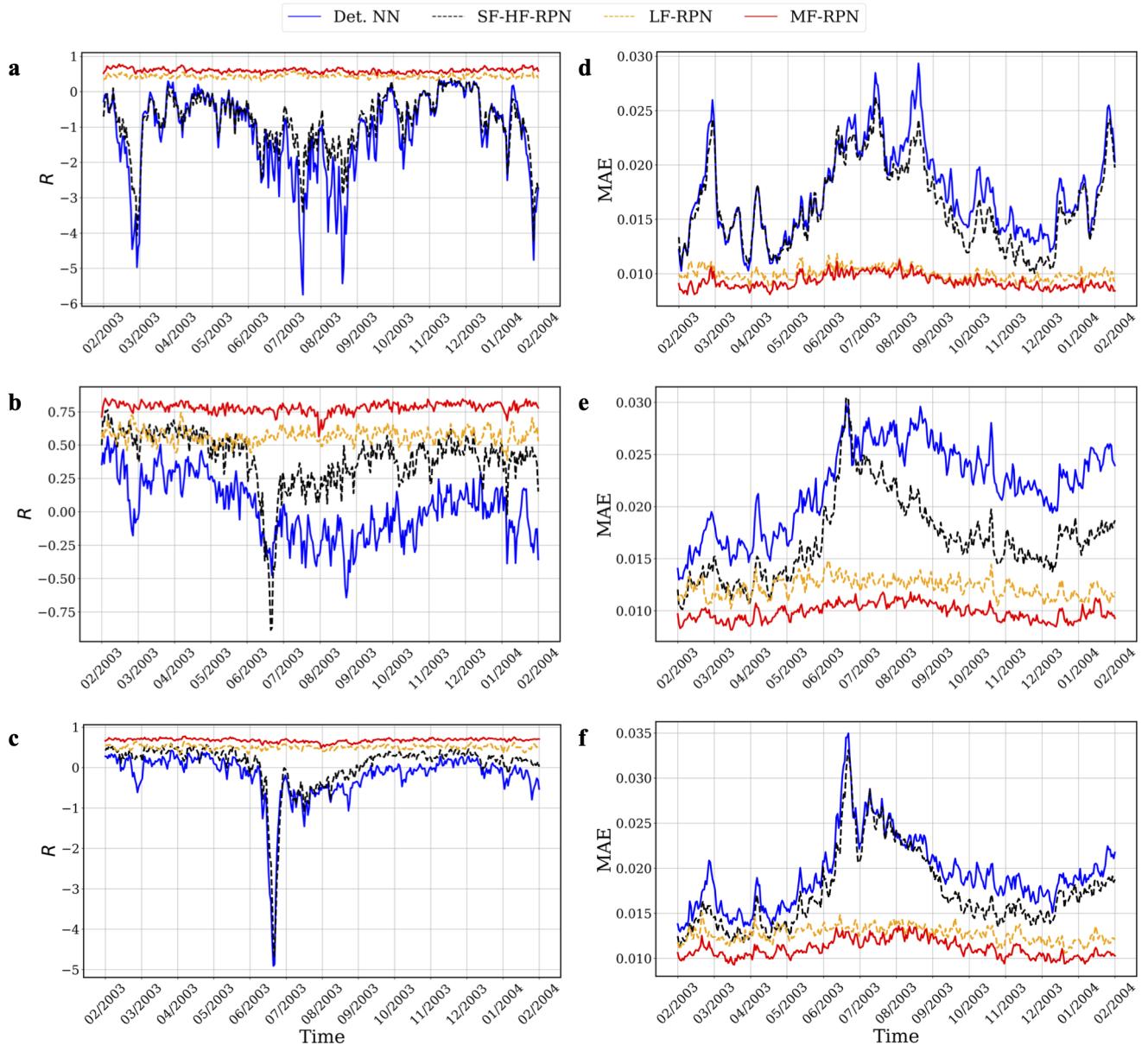
order to be able to have a longitude-latitude visualization. Hence the standard deviation considered always corresponds to the inherent one returned by the model and is not estimated by accounting for the temporal variance.

For the heat tendency, the longitude-latitude structure of the uncertainties returned by LF-RPN and MF-RPN follows relatively well the MAE variation in longitude-latitude directions with higher values around the tropics for both metrics and also occurring in the same regions (figures 13.a - 13.c and 14.a - 14.c). The corresponding uncertainty variations are nearly duplicates of the MAE longitude-latitude structures. However, we can clearly notice deviations in the uncertainty's longitude-latitude structure compared to the MAE variations for the SF-HF-RPN (figures 15.a - 15.c). Indeed, the latter tends to overestimate the uncertainty in the tropics and temperate zone for the vertical level 259 hPa (figure 15.a), while it underestimates the uncertainty in the same regions for the vertical level 761 hPa (figure 15.c).

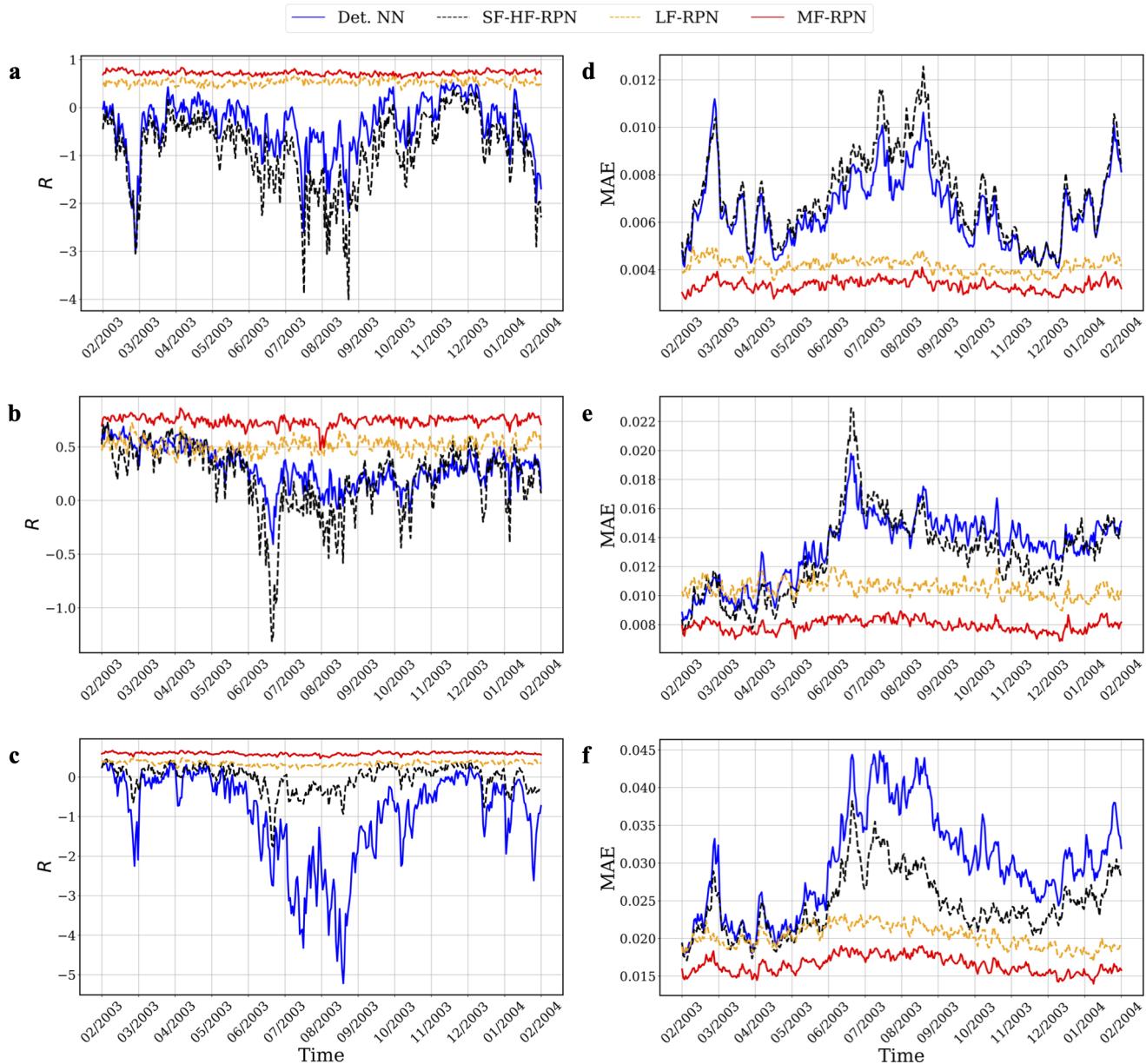
For the moisture tendency, there is also a good agreement between the longitude-latitude structures of the uncertainty and of the MAE returned by the MF-RPN and LF-RPN models for the vertical level at 259 hPa (figures 13.d and 14.d). For the moisture tendency at vertical levels 494 and 761 hPa, we still have some good agreement between the uncertainty and the MAE variations while more significant discrepancies are observed mostly within the tropics and temperate zone (figures 13.e - 13.f and 14.e - 14.f). The SF-HF-RPN's uncertainty also shows more important mismatches for the moisture tendency compared to the model's MAE for the vertical levels 494 and 761 hPa compared to the 259 hPa level (figures 15.d - 15.f).

Investigating the instantaneous variations (at the hourly time-step defining the testing data) of the MAE and returned uncertainty by the MF-RPN model also shows a nearly perfect agreement (figure 16). For instance, on February 1st 2003 at 14:00, all regions of high MAE values (south Atlantic ocean; south America; Pacific ocean; Australia and tropics within Africa and Indian ocean) are nearly perfectly replicated by the returned uncertainty  $\sigma_M$  by MF-RPN (figure 16.a). The same property is also observed at other time-steps, for instance on February 6th 2003 at 10:00, with the regions of high MAE values being different compared to the previous time-step, and the regions of high uncertainty  $\sigma_M$  shifting accordingly to well capture the regions of high MAE values (figure 16.b). Simulations of complete spatio-temporal evolution of MAEs and returned uncertainties for the heat and moisture tendencies by different models (MF-RPN, LF-RPN adn SF-HF-RPN) at vertical levels 259, 494 and 761 hPa can be found in [https://drive.google.com/drive/folders/1-RVxI\\_YpES0zNi6oa-0wIh8N1i6qL74Y?usp=drive\\_link](https://drive.google.com/drive/folders/1-RVxI_YpES0zNi6oa-0wIh8N1i6qL74Y?usp=drive_link). In addition to the instantaneous variations, the simulations include evolution of daily averaged predictions and uncertainties. All simulations show the ability of the MF-RPN to return uncertainties that nearly perfectly replicate the corresponding longitude-latitude error structure at all time instances and spatial points.

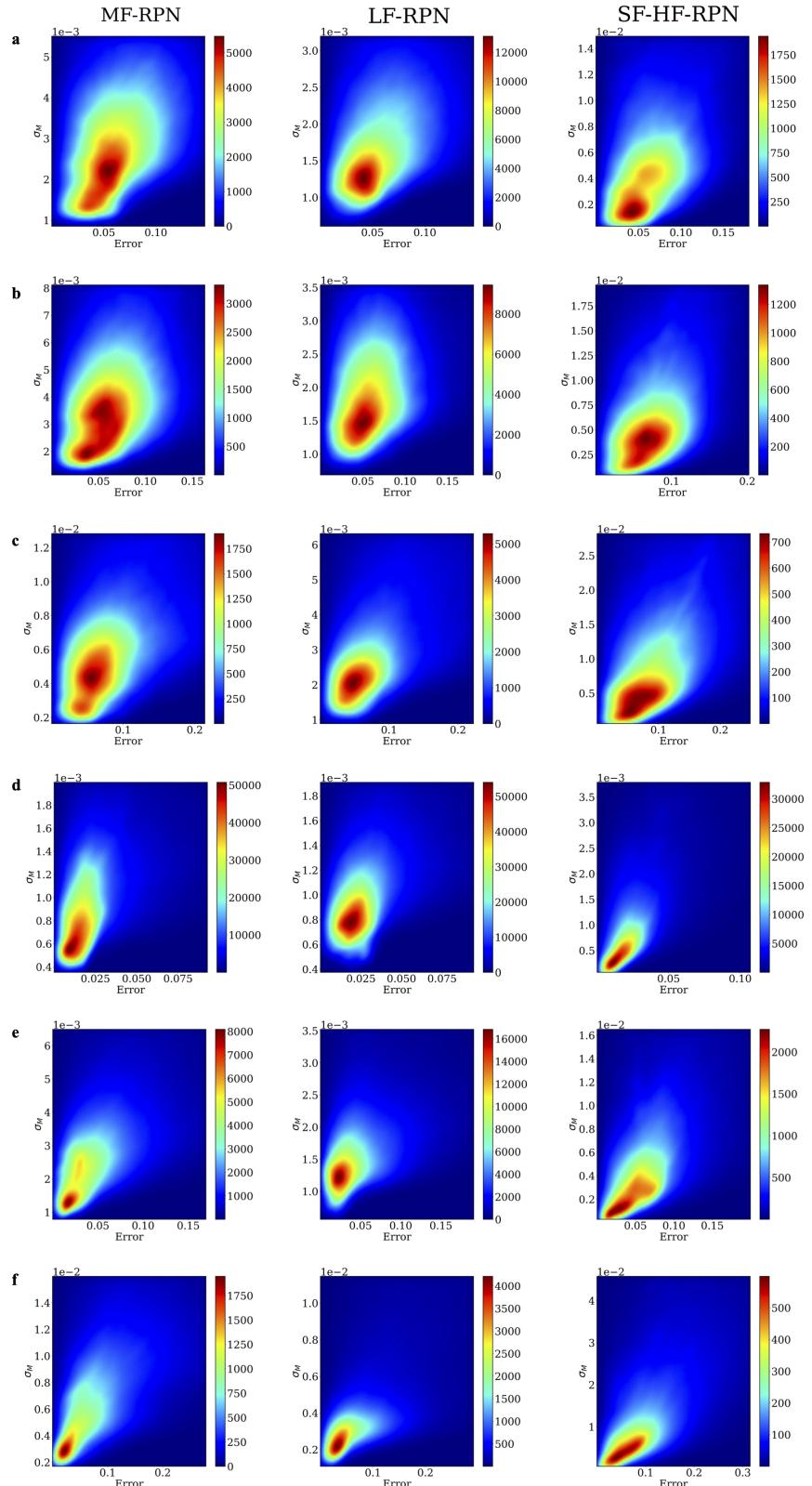
Overall, the MF-RPN model returns more trustworthy uncertainty quantification compared to the LF-RPN and SF-HF-RPN models. Hence, aggregating different datasets with different fidelity levels and of different climates through a well-designed multi-fidelity surrogate model does not only improve the parameterization's deterministic and statistical error metrics, but also provides more accurate uncertainty estimates. It is worth emphasizing again that the quantified uncertainty  $\sigma_M$  is estimated only based on the parameterization input without any knowledge of the true target value. Therefore, an accurate uncertainty quantification is a powerful tool, which can inform on how accurate the surrogate model is. This characteristic can be seen as similar to *a-posteriori* error estimates in model order reduction techniques.



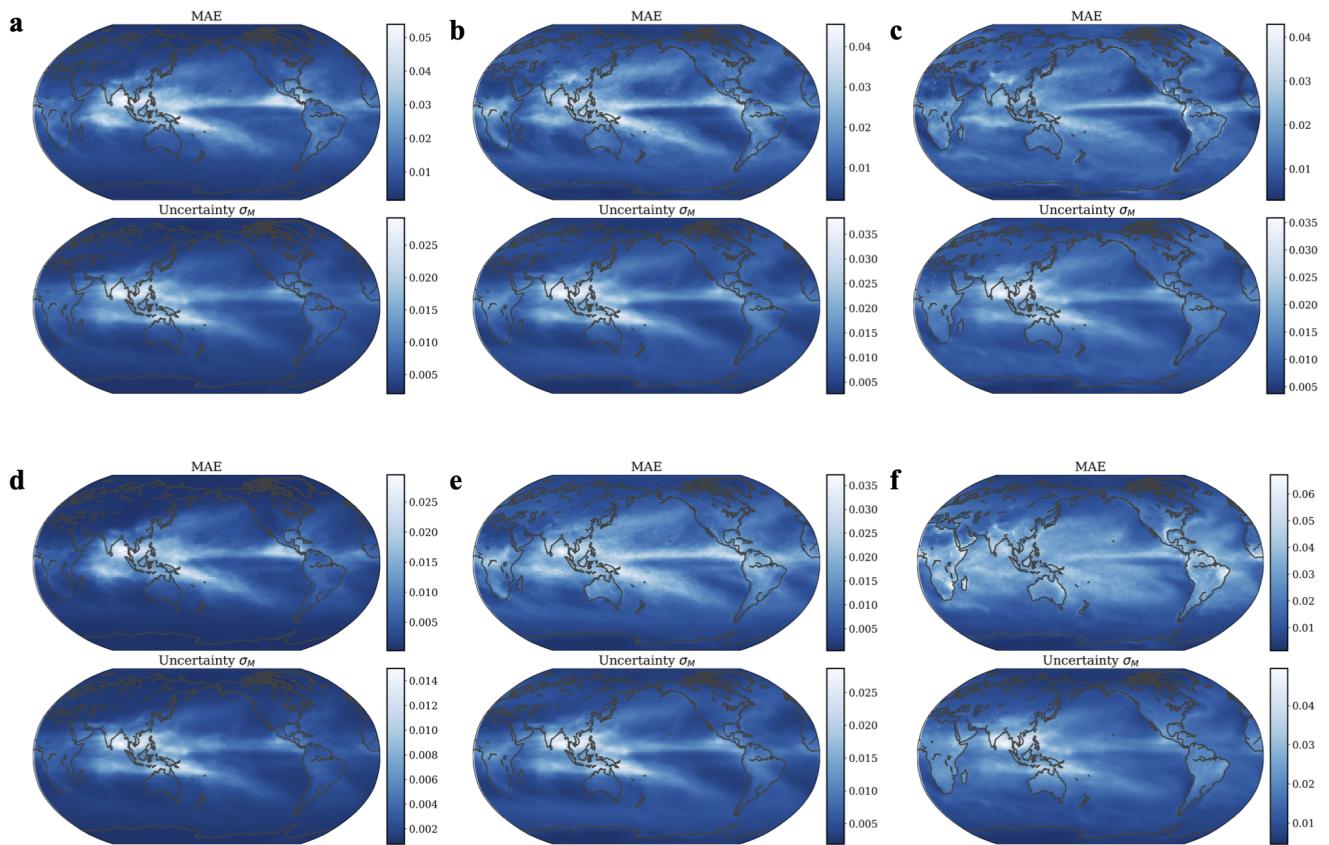
**Figure 10.** Temporal variation of MAE and  $R^2$  for heat tendency at different vertical levels and evaluated on test dataset. **a**,  $R^2$  at  $P = 259$  hPa. **b**,  $R^2$  at  $P = 494$  hPa. **c**,  $R^2$  at  $P = 761$  hPa. **d**, MAE at  $P = 259$  hPa. **e**, MAE at  $P = 494$  hPa. **f**, MAE at  $P = 761$  hPa.



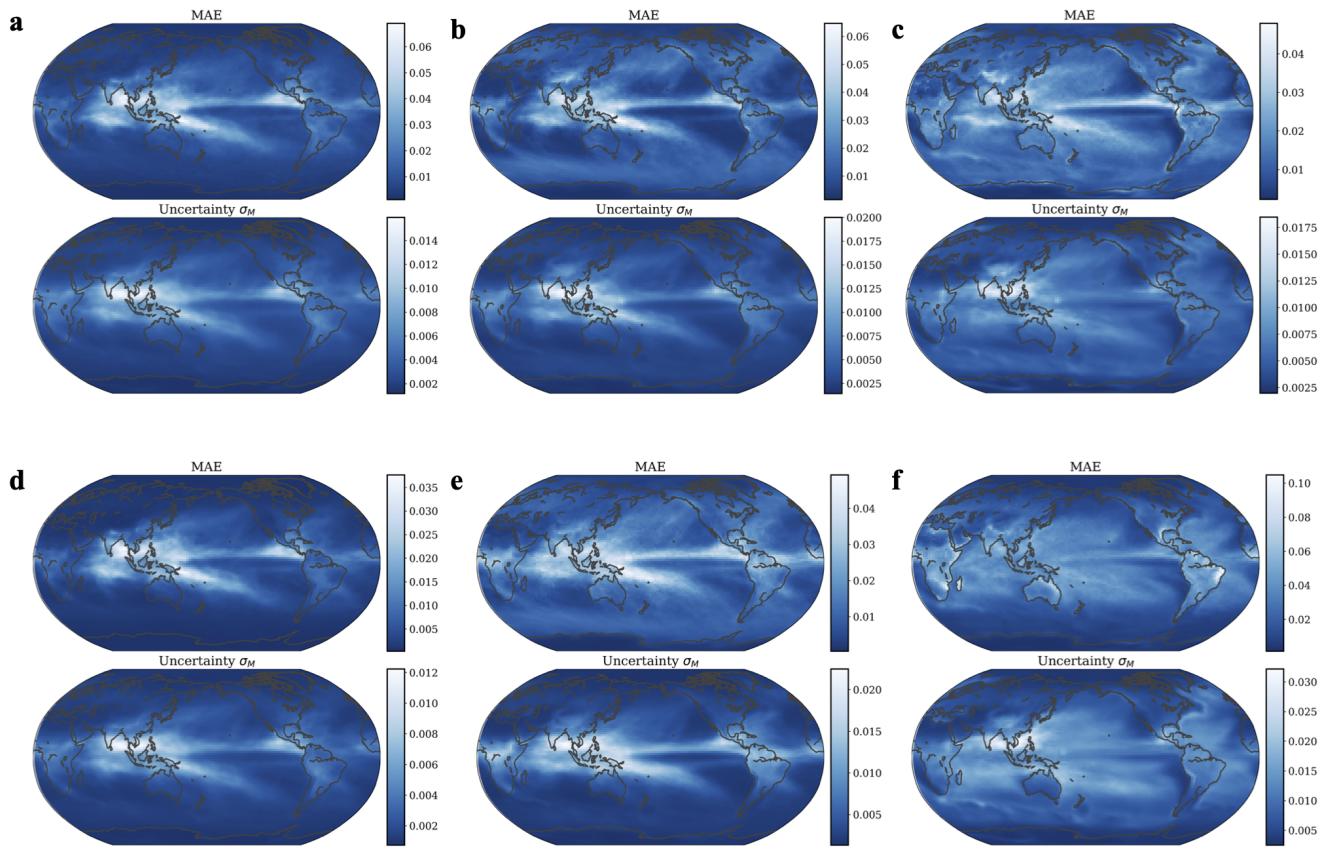
**Figure 11.** Temporal variation of MAE and  $R^2$  for moisture tendency at different vertical levels and evaluated on test dataset. **a**,  $R^2$  at  $P = 259$  hPa. **b**,  $R^2$  at  $P = 494$  hPa. **c**,  $R^2$  at  $P = 761$  hPa. **d**, MAE at  $P = 259$  hPa. **e**, MAE at  $P = 494$  hPa. **f**, MAE at  $P = 761$  hPa.



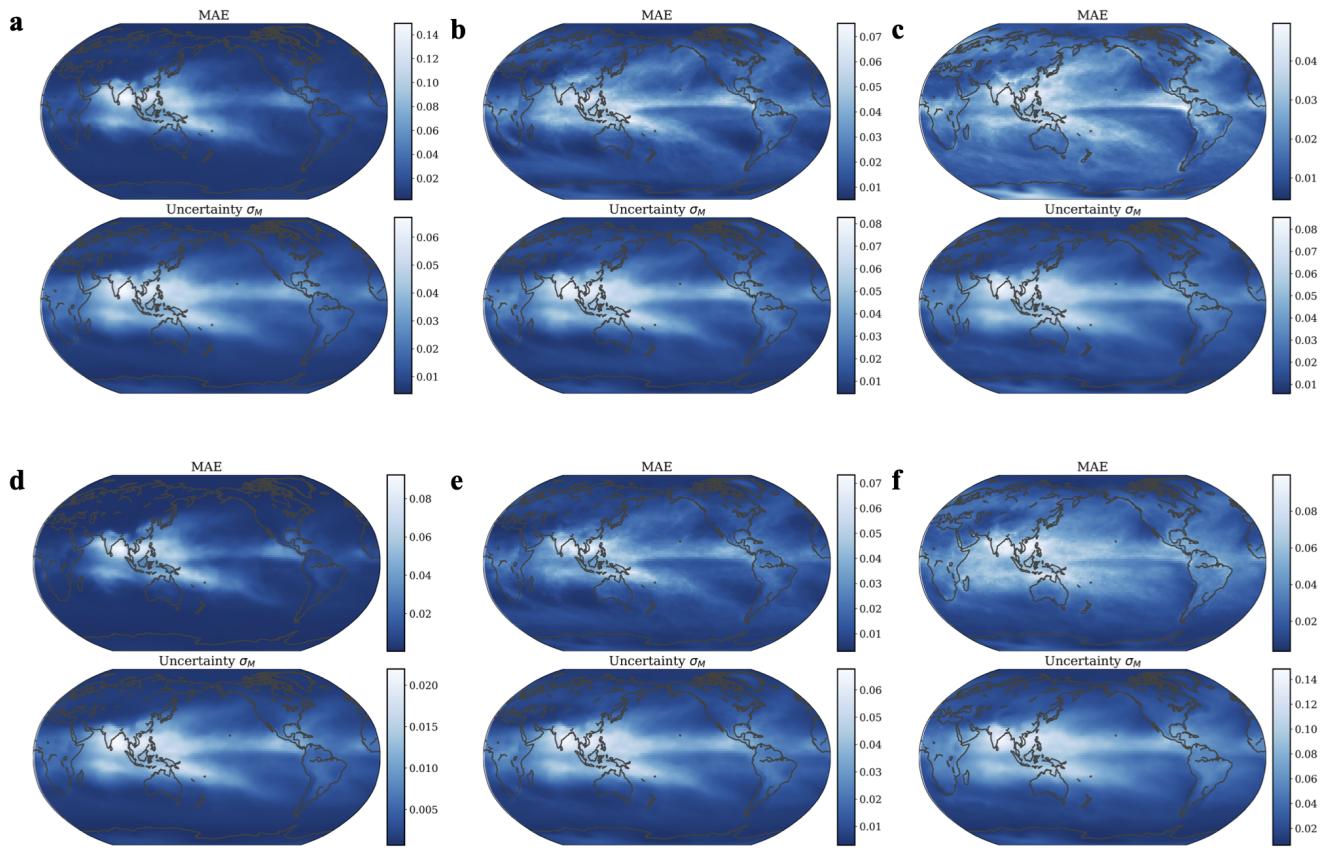
**Figure 12. Density plot of uncertainty  $\sigma_M$  as a function of error for testing points concatenated over space and time for MF-RPN, LF-RPN and SF-HF-RPN, heat and moisture tendencies and at different vertical levels. a, Heat tendency at  $P = 259$  hPa. b, Heat tendency at  $P = 494$  hPa. c, Heat tendency at  $P = 761$  hPa. d, Moisture tendency at  $P = 259$  hPa. e, Moisture tendency at  $P = 494$  hPa. f, Moisture tendency at  $P = 761$  hPa.**



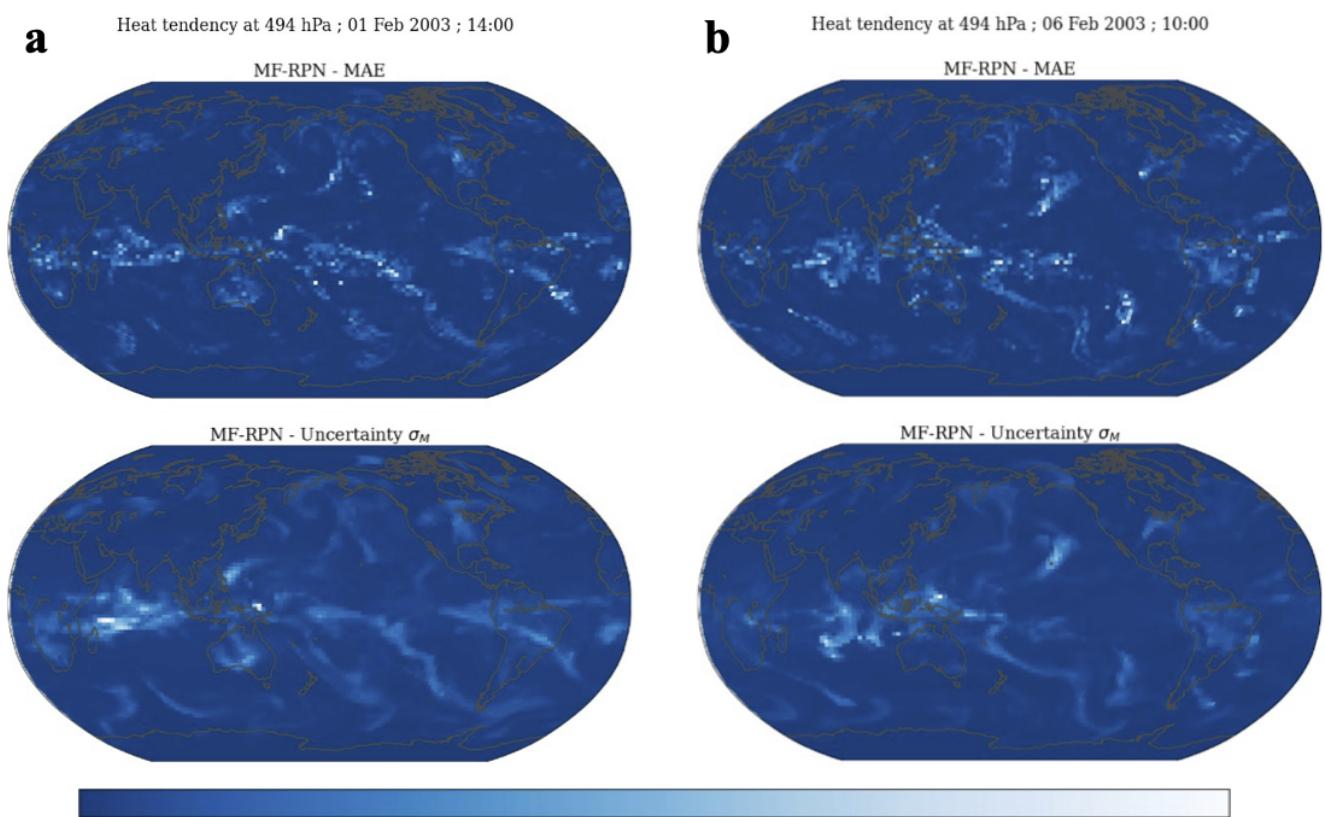
**Figure 13. Longitude-latitude variation of testing MAE and quantified uncertainty  $\sigma_M$  for MF-RPN for heat and moisture tendencies at different vertical levels.** **a**, Heat tendency at  $P = 259$  hPa. **b**, Heat tendency at  $P = 494$  hPa. **c**, Heat tendency at  $P = 761$  hPa. **d**, Moisture tendency at  $P = 259$  hPa. **e**, Moisture tendency at  $P = 494$  hPa. **f**, Moisture tendency at  $P = 761$  hPa.



**Figure 14. Longitude-latitude variation of testing MAE and quantified uncertainty  $\sigma_M$  for LF-RPN for heat and moisture tendencies at different vertical levels.** **a**, Heat tendency at  $P = 259$  hPa. **b**, Heat tendency at  $P = 494$  hPa. **c**, Heat tendency at  $P = 761$  hPa. **d**, Moisture tendency at  $P = 259$  hPa. **e**, Moisture tendency at  $P = 494$  hPa. **f**, Moisture tendency at  $P = 761$  hPa.



**Figure 15. Longitude-latitude variation of testing MAE and quantified uncertainty  $\sigma_M$  for SF-HF-RPN for heat and moisture tendencies at different vertical levels. **a**, Heat tendency at  $P = 259$  hPa. **b**, Heat tendency at  $P = 494$  hPa. **c**, Heat tendency at  $P = 761$  hPa. **d**, Moisture tendency at  $P = 259$  hPa. **e**, Moisture tendency at  $P = 494$  hPa. **f**, Moisture tendency at  $P = 761$  hPa.**



**Figure 16.** Longitude-latitude variation of testing MAE and quantified uncertainty  $\sigma_M$  at given time instances for MF-RPN for heat tendency at vertical level  $P = 494$  hPa. **a**, on Feb 1st 2003 at 14:00. **b**, on Feb 6th 2003 at 10:00.