Response to Editor

In the following document, the editor's comment is shown in **black** and our response is in **blue**.

Dear Authors,
The reviewers are generally pleased with the improvements to your manuscript, but one reviewer still has some concerns that should be addressed.

In the revised version of the paper, we have addressed the reviewers' comments. We hope our responses and the modifications proposed will be suitable for them.

We note that the two reviewers do not agree on the content of the article. The reviewer 2 accepts the paper as it is but suggests that the addition of the realistic section (which was requested by the reviewer 3) could benefit from another separate paper dedicated to a more detailed study. On the contrary, the reviewer 3 regrets the absence of online tests.

We agree with the reviewer 2 that the paper as it stands, is self-contained. For this reason, we do not consider it would be useful to add a new section that focuses on the online evaluation. This crucial step leads to a complete exploration that could be further developed in a future work and may be the subject of a new paper.

Response to Referee #2

We would like to thank the reviewer 2 for taking the time to review this article.

In the following document, the reviewer's comments are shown in **black** and our point-by-point responses are in **blue**.

I would like to thank the authors for carefully considering my comments and addressing them thoroughly. I find that the revisions, including the incorporation of the reviewers' feedback and the decision to move part of the figures to the Supporting Information, have improved the clarity and overall quality of the manuscript.

However, I am somewhat intrigued by the authors' decision to add a section with preliminary results on applying the method in a realistic geometry setting (Section 6), at the end of the manuscript (although I understand that this addition was suggested by Reviewer #3). While these results appear valuable and interesting, I believe they are quite preliminary and would benefit from further investigation if possible. In my view, they may be better suited for a separate, dedicated publication once more comprehensive analyses are available.

We totally agree, this could be the subject of a full study. However, through these preliminary results, we show the potential of the emulators on a realistic setup that it is relevant and interesting to present in this article in order to motivate further and in-depth analyses in the future.

Response to Referee #3

We would like to thank the reviewer 3 for taking the time to review this article and for helpful comments.

In the following document, the reviewer's comments are shown in **black** and our point-by-point responses are in **blue**.

The manuscript presents an approach that integrates latent physical knowledge into the physics emulator and, in the revised version, expands the data set to include a more realistic geographical configuration. This addition is a clear improvement and demonstrates the authors commitment to enhancing the relevance of their experiments.

Nevertheless, the most critical shortcoming that remains unaddressed is the absence of online (stability) testing of the proposed schemes. While offline performance metrics are informative, they do not guarantee that the emulator will remain numerically stable when coupled to the full model and operated in a prognostic setting.

As discussed previously, we fully agree with the reviewer that the online tests are important to investigate the accuracy and the stability of simulations with the emulator. However, this is left for future work, which may be the subject of a new paper focusing on the evaluation of the performance of the emulator in this type of online setup. Here, in this first study, we aim to highlight how emulators, without being coupled with the dynamical core, reproduce the behaviour of the physics and how the addition of physical knowledge allows results to be closer to those from the physical model. This offline study is necessary before proceeding to the online evaluation to ultimately adopt the emulator as a substitute for traditional physical parameterizations. Therefore, we do not wish to add a section on the online evaluation to this article. Moreover, the reviewer 2 has already suggested that the realistic study could benefit from an entire paper but accepts the paper as it is.

Below I outline several major points that require further attention, followed by more minor comments.

 # Major comments

1. Line 483: "We consider δz = 1. This distance is fixed as the step between levels is constant." LMDZ uses hybrid sigma-pressure coordinate in which the distance between adjacent levels is not fixed to my knowledge

   Indeed LMDZ uses a hybrid pressure-based coordinate, and the exact expression of the tendency due to turbulence depends on the variable

spacing between levels. We have removed some contents and provided a corrected explanation in the paper:

"*The phyLMDZ model uses a hybrid pressure-based coordinate. Therefore, the thickness (in m) and mass content (in kg m^{-2}) of each model layer vary with model level. Both affect the precise expression of the tendencies due to turbulence. For the zonal velocity at the level l, this expression is of the form:*

$$\frac{\partial u_l}{\partial t} = K_l^+\left(u_{l+1} - u_l\right) - K_l^-\left(u_l - u_{l-1}\right)$$

where the coefficients K_l^+ and K_l^- incorporate the turbulent diffusivities K_z computed above and below level l, the mass content of layers and their thickness. In order to reveal latent structural information without exposing too many details of the underlying parameterization, we use the above formula with $K_l^+ = K_l^- = 1$, i.e. a simple discrete three-point laplacian Δu_l, as additional input".

An advantage of this discrete vertical laplacian formulation for the sake of comparing the emulators, is that the laplacian computed in this way can also be computed by a convolutional filter in the U-Net architecture.

2. You state that the DNN has six times less the number of parameter than the U-Net. Could the improved performance of the U-Net compared to DNN be simply due to the additional complexity/expressivity of the U-Net?

   The U-Net we used has indeed six times more parameters than the DNN we used. It is more flexible in fitting the relationship between the inputs and the outputs. The main challenge in deep learning or with large models with millions of parameters (the DNN has a couple of millions of parameters) is to ensure that the models do not overfit the training data. In our case we use early stopping and the best models have been optimised only on a relatively low number of epochs. The evaluation of the model has been made on a distinct time period to minimize the correlation between the training and test datasets. The higher number of parameters can play a role on why the U-Net performs better than the DNN. We have tested that hypothesis by constructing a DNN with a number of parameters close to that of the U-Net (around 13 million parameters). This DNN has 12 hidden layers of 1,024 neurons each. We have trained it several times and the conclusion is the same for each model: increasing the number of parameters does not improve the performance of the DNN, quite the contrary, there is a deterioration in terms of loss function as shown in Table 1.

	U-Net		Initial DNN		New DNN	
	with laplacians	without laplacian	with laplacians	without laplacian	with laplacians	without laplacian
training loss score	2.32×10^{-1}	3.26×10^{-1}	2.86×10^{-1}	4.24×10^{-1}	3.43×10^{-1}	5.24×10^{-1}
validation loss score	3.64×10^{-1}	3.94×10^{-1}	4.24×10^{-1}	4.96×10^{-1}	5.03×10^{-1}	5.64×10^{-1}

Table 1. Training and validation loss function scores for the U-Net, the initial DNN (the one used in the article) and a new DNN (the one with a similar number of parameters as the U-Net). Only one DNN model is shown in this table.

We believe that the main reason the U-Net performs better is because it is a fully convolutional neural network with a U-shape structure, which allows it to extract spatial information in the column at multiple scales from the input and to enforce spatial structures in the outputs. A DNN with enough (i.e., many more) neurons could potentially do the same but the parameter space to explore would be much larger. For instance, to compute the laplacian as implemented in the paper after the input layer, a U-Net only needs to find the 3 parameters values of a 3x1 filter. Reproducing such a filter in a DNN would involve many more neurons as neurons between layers are fully connected and would lead to a much more complicated optimization problem with several local minima.

 # Minor comments

1. Figure 2: Why is the number of cells for the test set the same as for the other sets? I thought you only apply temporal sampling for the test set - is that wrongly visualized?

   Thanks for pointing this out, this was a mistake. Consequently, this figure has been modified accordingly. The paragraph on spatial sub-sampling has also been modified as follows:

   "*We then perform temporal and spatial sub-sampling of the datasets to avoid spatio-temporal correlations and to reduce the cost of training. For this purpose, starting from the first time step for each dataset, we select one data point every ten time steps and thus obtain datasets with 2,074 time steps for the training dataset and 692 time steps for the validation and test datasets. Moreover, we choose to take one cell every twenty cells to create the final training and validation datasets. In this way, we obtain data that is well distributed across the globe and we ensure that we have sampled at all latitudes (see Fig. S1 in the Supplementary Material). Finally, the training*

dataset contains data from 801 cells and the validation dataset has 800 cells spread across the globe among the 16,002 cells. We took care to ensure that this spatial selection represented all the regions. Therefore, we have 1,661,274 samples (2,074 x 801 grid points) for the training dataset. We validate our models on a dataset with 553,600 samples (692 x 800 grid points). In order to study the results of the emulators across all grid cells of the globe, the temporal sub-sampling has not been applied for the test dataset".

2. Line 267: But the stopping criterion is applied for the validation loss, right?

   Exactly, the early stopping criterion is applied to the validation loss. We have modified the sentence as follows:

   "*Early stopping criterion has been used to prevent overfitting since this option monitors the performance of the model on a validation dataset during the training process. Thus, training is stopped when the performance on the validation loss no longer improves*".

3. To my previous comment on Line 483: I don't see why the multivariate nature of the problem should prevent you from showing multiple training/validation loss curves, but at least you added the information that you observed similar convergence behavior for multiple training runs.

   We are happy that the added information suits the reviewer.

4. Line 436: "... gravity wave drags ... due to convective precipitation ..." please change "precipitation" to "systems" or explain the mechanism you have in mind.

   As suggested by the reviewer, the word "precipitation" has been changed by "systems" in the revised version of the article.

5. Line 536ff: Your argumentation seems flawed here. If turbulence is not involved in the computation of d_qx3, the DNN with added Laplacians as inputs should not be able to outperform the U-Net without Laplacians, right?

   Thank you for pointing this out. We have removed the sentences that were about the involvement of turbulence in the ice water tracer tendency.

6. Figure 6: The inset have no labels.

   In the revised version of the article, we have modified the figures where there are inserts (figures 5b, 6e, 6f, 7e, 7f, 8e and 8f) to add the x- and y-axes tick labels.

8. Discussion of degradation in skill to represent variance in output profiles with the U-Net when adding Laplacian as input is missing.

   Indeed, we did not include a detailed discussion about the degradation of the global variance profile of the U-Net when we add laplacians for the realistic

setup. We do not have a clear explanation for this. In-depth analyses are needed to fully understand this degradation, However, in the following, we provide a few comments to add to the discussion.

We have computed the global RMSE and the global R^2 of the zonal wind tendency for the test subset, for each emulator. The results are presented in Table 2 below. Based on these results, we can highlight the difference between emulators in terms of performance. Despite the fact that there is a degradation of the global variance profile of the zonal wind tendency, the U-Net with laplacians achieves the highest scores and performs better than the one without these variables, in terms of global metrics.

NN-model	RMSE (m/s^2)	R^2
DNN, initial training	1.96×10^{-4}	0.42
U-Net, initial training	1.62×10^{-4}	0.60
DNN, training with Δ	1.76×10^{-4}	0.53
U-Net, training with Δ	**1.56×10^{-4}**	**0.63**

Table 2. Results of the general metrics for the zonal wind tendency over all time steps, cells and vertical levels of the test subset, to evaluate the four emulators studied.

We have also computed the mean bias and the variance ratio of the zonal wind tendency at the 3rd vertical level, where the variability is strongest (see Fig. 11 of the article). The results are illustrated below with Figure 1 and Figure 2, respectively. In general, for this vertical level, we can underline that emulators show bad performances in terms of mean bias where the orography is important. The variability deteriorates at well-defined spatial locations and, as mentioned before, further analyses are needed to better understand this degradation.

We have added the following sentences in the revised version of the paper:

"*It would appear that this deterioration occurs in favor of the representation of the global mean profile. The U-Net with laplacians still remains the best model overall in terms of global RMSE for the zonal wind tendency. Indeed, this global RMSE is equal to 1.56 x 10^{-4} m s^{-2} for the U-Net with laplacians, compared to 1.62 x 10^{-4} m s^{-2} for the U-Net without laplacian, 1.76 x 10^{-4} m s^{-2} for the DNN with laplacians and it reaches 1.96 x 10^{-4} m s^{-2} for the DNN without laplacian. The addition of orography and continents can lead to significant differences in terms of the intensity of physical processes that are certainly linked to the spatially varying performances of the emulators in retrieving the mean and variance profiles of the zonal wind (not shown). With*

the addition of orography, the discrete approximation of the laplacian used as input may not represent as well the effect of turbulence on the zonal wind vertical structure. Further analyses need to be conducted to clearly understand this degradation".

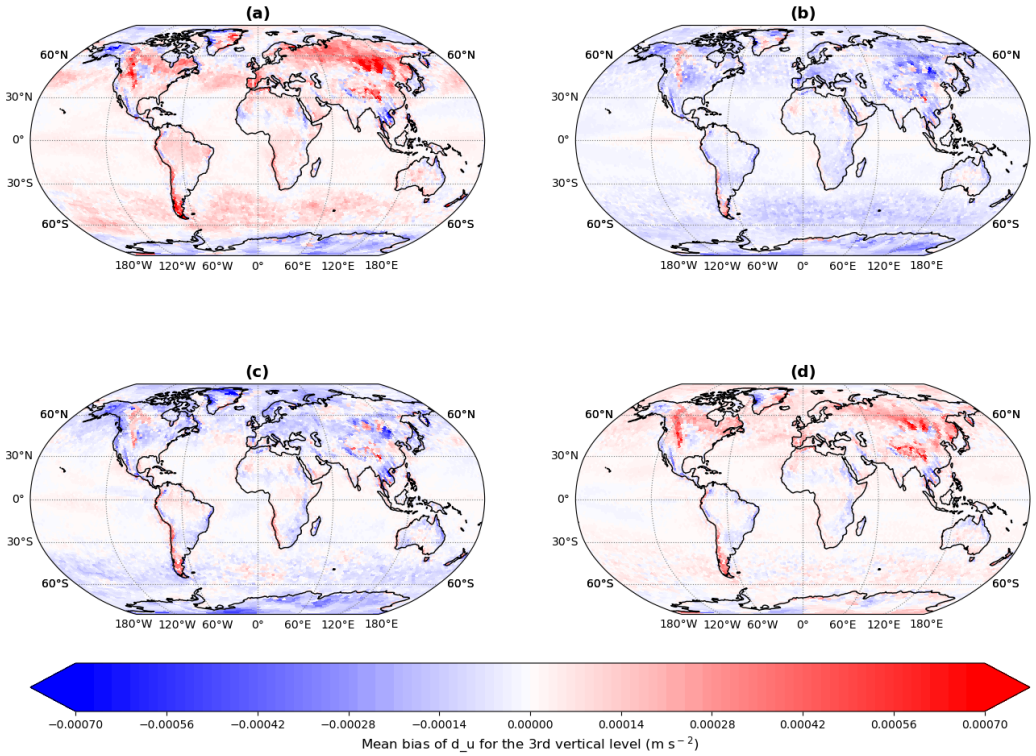

Figure 1. Map of the mean bias of the zonal wind tendency d_u at the 3rd vertical level for the test subset with the realistic configuration, calculated between data from ICOLMDZ and predictions made by an emulator. Figure (a) corresponds to the results from the DNN without laplacian, (b) those from the U-Net without laplacian, (c) those from DNN with laplacians and (d) those from the U-Net with laplacians.

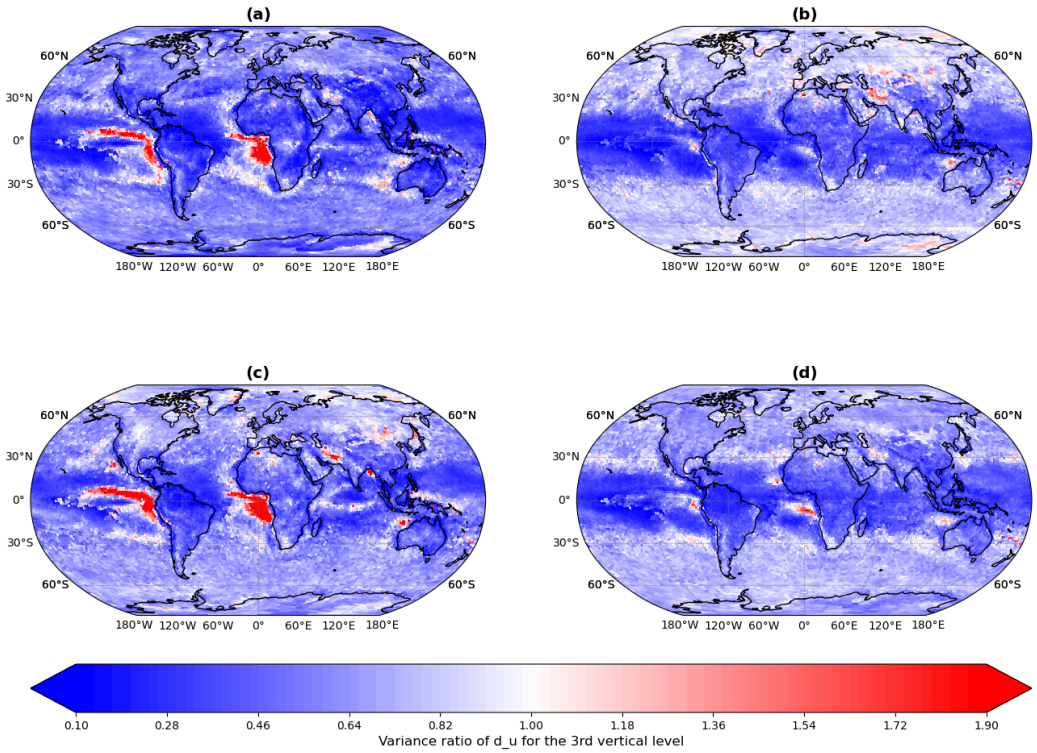

Figure 2. Maps of the variance ratio of the zonal wind tendency d_u at the 3rd vertical level for the test subset with the realistic configuration, calculated between data from ICOLMDZ and predictions made by an emulator. Figure (a) corresponds to the results from the DNN without laplacian, (b) those from the U-Net without laplacian, (c) those from DNN with laplacians and (d) those from the U-Net with laplacians.

 # Typos/Grammar

1. gravity wave drags -> gravity wave drag

   This has been corrected.

2. Line 600: "... where topography is marked". Some other word like "heterogenous" instead of "marked" would fit better here.

   This has been corrected with the word "heterogeneous" as suggested by the reviewer.