Response to reviewer

R. El Montassir, O. Pannekoucke and C. Lapeyre

June 4, 2024

We thank the reviewer, Dr. Alban Farchi, for his feedback on our manuscript. We appreciate the positive remarks and constructive suggestions. We are grateful for the time and effort that he invested in evaluating our work.

We copied the reviewer's commentary below, and we replied in teal to each point. We also provide the changes made in the manuscript for each comment.

1 General comments

1.1 Objective of the manuscript

After reading the entire manuscript, I am not entirely sure what is the objective. The title suggest that the objective is to develop and release a version of a geoscientific model, namely HyPhAI. On the other hand the last paragraphs of the introduction, and more generally the entire manuscript, leave the impression that the authors want to develop a method and illustrate it to a specific application (cloud cover nowcasting).

Finally, even though the name 'HyPhAI' corresponds to the model that is described here (which is indeed an hybrid model mixing physics and AI), this name could fit any hybrid model mixing physics and AI and hence it gives the false impression that the model describes all possible solutions for hybridising physics and AI. Therefore, I would suggest to be slightly more humble in choosing the name of the model.

- Regarding the objective of this work, it is to develop a hybrid architecture that combines physics and AI that can be applied to a wide range of geoscientific problems. The cloud cover nowcasting is just an application of this architecture. We agree that the title was misleading, and we changed it to "A Hybrid Physics-AI (HyPhAI) approach for probability fields advection: Application to cloud cover nowcasting". We also changed the names of the models to HyPhAICCast-1, HyPhAICCast-2, HyPhAICCast-3, and HyPhAICCast-4.

1.2 Background in machine learning and numerical schemes for PDEs

A large part of section 2 is widely known in the community and, if needed by a reader, can be found in standard textbooks. In particular, I would suggest to remove

- *in subsection 2.2 the last two paragraphs, from L 124 'During this training process...' to L 136 '... employed in neural networks.';*
- *the entire subsection 2.3;*
- *the entire subsection 2.4;*
- *the entire subsection 2.5;*

or at the very least to put all these materials in appendices.

- Similar comments were made by the other reviewer. We are aware that the section contains standard information; however, we do not fully agree that these details are known to the majority of the readers, especially in our restricted community, meteorology. We believe that it is important to provide a clear and detailed explanation of the challenges encountered when combining neural networks with physics-based models, in order to provide a smooth reading experience for the reader. However, we have moved this Section 2 to the Appendix and updated the introduction of Section 3 as follows:

"In this ~~section, we introduce our hybrid~~ work, we address applications involving dynamics with unknown variables that require estimation. For example, the cloud motion field is one of the unknown

variables in the application considered. In such cases, as discussed in the Introduction, a joint resolution approach is more appropriate. Here, the physical model uses the neural network outputs to compute predictions, integrating the two models as follows:

$$y = \phi \circ f_\theta(x),$$

where x is the input, f_θ represents the neural network, ϕ denotes the physical model, and y is the output. In this setup, ϕ implicitly imposes a hard constraint on the outputs, potentially accelerating the convergence of the neural network during training.

This method raises some trainability challenges as the physics-based model is involved in the training process, and it should be differentiable, in the sense of automatic differentiation, in order to allow the back-propagation of gradients (refer to Appendix B). We show in Appendix B how spatial derivatives of PDEs can be approximated within a neural network in a differentiable way using convolution operations. This allows us to compute gradients and back-propagate them during the training process. This fundamental knowledge serves as a foundation for our investigation of novel hybrid Physics-AI architectures. With these established principles, we present in this section the proposed hybrid architecture, which is applied to cloud cover nowcasting. In this section, we introduce our hybrid Physics-AI architecture, ~~denoted as HYPHAI (an abbreviation for Hybrid Physics-AI),~~ detailed in Sect. ~~3.1. Section 3.2~~ 2.1. Section 2.2 explains the different physical modelling approaches investigated in this study. Following that, Sect. ~~3.3~~2.3, Sect. ~~3.4~~ 2.4 and Sect. ~~3.5~~ 2.5 sequentially present the training procedure, evaluation metrics, and benchmarking procedure."

1.3 Two or four model variants

In subsection 3.2, four model variants are presented, whereas in the numerical experiments, only the first two are tested. Therefore, I would highly recommend to remove HyPhAI-3 and -4 from here to simplify the presentation (which is already rather complex). Then HyPhAI-3 and -4 could be either presented in appendices or in the discussion.

- The same point was raised by the other reviewer. The reason why these two versions were presented even if they didn't show any improvement is to show the flexibility of the model. We added a sentence to clarify this, and we moved these two versions to the appendix.

"The second version of the hybrid model, denoted ~~HyPhAI~~HyPhAICCast-2, adds this source term to the advection. This modelling is described in the following equations:

$$\partial_t P_j + \vec{V} \cdot \vec{\nabla} P_j = \tanh(S_j) \quad \forall j \in \{1, 2, \ldots, C\}, \tag{1}$$

where S_j is estimated using a second U-Net model (see Fig. 4). While the previous modelling describes the missing physical process in the advection, it does not satisfy the probability conservation property. Thus, this modelling does not conserve the probabilistic nature of P over time. To ensure the appropriate dynamics of probability, a robust framework is provided by continuous-time Markov processes across finite states [Pavliotis and Stuart, 2008, chap. 5]. In this framework, the probability trend is controlled by a linear dynamics, ensuring the bound preservation, positivity, and probability conservation. Two other models based on this framework, named HyPhAICCast-3 and HyPhAICCast-4, are presented in the Appendix A1 and Appendix A2. However, these models did not show any performance improvement compared to the simpler HyPhAICCast-1. Indeed, beyond the performance aspect, this hybridisation framework is flexible, not only limited to the advection, and can be extended to other physical processes"

1.4 Application to the Earth's full disk

This inference experiment is very nice, but raises two key questions:

1. *What is exactly a 'full disk' and how is it projected into the* 3712×3712 *squared image?*

   - A satellite full disk view of the Earth is a single image taken by a satellite in geostationary orbit, capturing an entire hemisphere of the Earth's surface in one frame. The 3712×3712 image simply corresponds to the image captured by the satellite, in this case the satellite is a geostationary satellite called Meteosat Second Generation (MSG) provided by EUMETSAT and positioned at 0° longitude. Here is the modified sentence:

"..., we tested it on a much larger domain, ~~the full earth's disk~~ an entire hemisphere of the Earth - also called a full disk - centred at 0 degrees longitude. ~~This~~ The satellite observations of this expansive full disk domain ~~is 14 times~~ are of size 3712×3712, which is 210.25 times larger than the size of the training ~~area~~ones."

2. *Beyond the visual impressions from figures 12 and A3, do you have scores to support the claim of 'remarkable adaptation' (L 481) and 'accurate and reliable' (L482)?*

   - We agree that the visual impressions are not sufficient to support these claims, however it's irrelevant to compare scores over different domains. We rephrased the sentences:

   "..., we tested it on a much larger domain, ~~the full earth's disk~~ an entire hemisphere of the Earth - also called a full disk - centred at 0 degrees longitude. ~~This expansive full disk domain is 14 times~~ The satellite observations of this expansive full-disk domain are of size 3712×3712, which is 210.25 times larger than the size of the training ~~area~~ones. It has diverse meteorological conditions and includes projection deformations when mapped onto a two-dimensional plane~~. Therefore, it provides an ideal~~ , while the extreme deformations at the edge of the disk make this data less useful for operation purposes, it still provides an interesting testing ground for ~~HyPhAI-1~~HyPhAICCast-1's generalisation ability. In this analysis, we focus only on visual aspects. Despite the significant differences between the training domain and the full disk, we observed ~~a remarkable adaptation of the HyPhAI-1 model to this new context~~ good qualitative forecasts of the HyPhAICCast-1 model on this new domain without any specific training on it (see Fig. ~~A3~~12 and Fig. A4). The cloud motion estimation on the full disk was found to be ~~accurate and reliable, this~~ visually consistent, a video supplement is provided in the supplementary material. This successful transferability of the model highlights its potential robustness ..."

2 Technical comments and suggestions

L 17-18 *'However, NWP models have inherent limitations in their ability to capture small-scale weather phenomena such as thunderstorms, tornadoes, and localised heavy rainfall events.' Please add a citation here.*

- We added the following references: [Schultz et al., 2021, Matte et al., 2022, Joe et al., 2022].

L 33 *'SHI et al., 2015' Is the capitalisation of the name intentional?*

- Corrected.

L 37 *'This network excels' I would rather speak of 'neural architecture' as LSTM is no unique network.*

- Corrected.

L 52 *'the hybridisation available techniques'* $\rightarrow$ *'the available hybridisation techniques'?*

- Done.

L 52-53 *'As discussed by Willard et al. (2022), the hybridisation available techniques leverage different aspects of ML models, e.g. the cost function, the design of the architecture or the weights' initialisation.' I would suggest to also cite here the review by [Cheng et al., 2023].*

- Done.

L 68-69 *'However, it does not have the ability to enforce physics-based constraints, as it primarily deals with errors rather than physical states.' Why not? Even in residual modelling, nothing prevents you from adding enforce a physics-based constraint.*

- Yes, but it would not be in this category of methods, either in the first one (L 55-60) or in the next one. However, this sentence is misleading and not necessary, we removed it.

L 70 *'An advanced variation of residual modelling involves the integration of physics based models and ML models.' I am not sure to see the difference here. In what you call residual modelling, the ML model predicts the errors of the physics-based model, in such a way that the final model is hybrid and aggregates the contribution of the physics-based model and of the ML model, which is precisely what you describe in the second part of this sentence.*

- Residual modelling is a specific case, and what we describe in the following lines is more general. We rephrased the sentence to make it clearer:

"To address imperfections in physics-based models, a common strategy is ~~residual~~ error modelling. Here, an ML model learns to predict the errors (also called residuals) made by the physics-based model (Forssell and Lindskog, 1997). This approach leverages learned biases to correct predictions (see Fig. 1.). ~~However, it does not have the ability to enforce physics-based constraints, as it primarily deals with errors rather than physical states.~~

~~An advanced variation of residual modelling involves the integration of~~ A more general approach that does not deal only with errors is to create hybrid models merging physics-based models and ML models. ~~In scenar- ios~~ For example, in scenarios where the dynamics of ~~Physics are~~ physics is fully defined, ~~a straightforward method involves using~~ the output of a physics-based model can be used as an input to an ML model."

Equation 4 *The notation $f_\theta(x_k)$ here is inconsistent with the notation $f_\theta(x, x_{Phy})$ in Eq. 3 .*

- We thank the reviewer for pointing this out, we replaced $\phi \circ f_\theta(x_k)$ by $f_\theta(x_k)$.

L 122-123 *'The choice of l depends, among other things, on the statistical model f_θ.' Here I disagree. The choice of the likelihood should not depend on the model, but it should be the other way around: the choice of the model should be made in order to be able to minimise the likelihood.*

- We thank the reviewer for pointing this confusing sentence out. What we meant is for example, in image generation, the MSE loss can be used for a DDPM model, while GAN require a completely different loss function. But we agree that this could be seen in the other way around, and that "statistical" is just adding confusion, we removed this sentence.

Figures 3 and 4 *Is this a game of 'find 7 differences'? On a more serious note, I wonder whether these two figures could be merged.* - We don't see how these two figures could be merged, each one shows a different model. The difference between the two figures is that Figure 4 adds a source term to the equation, and this source term is estimated by a U-Net, this is explained in the caption and also in the text (subsection 3.2.2). However, we have reduced the opacity of the unchanged parts in the second diagram and highlighted the additional parts.

L 234 *'These 256×256 satellite images' I assume that 256×256 is the size of each image, but here one could naively understand that there are 65536 images in total.*

- We rephrased the sentence:

"...~~and~~, the time step is 15 minutes ~~. These 256×256 satellite~~ and each image is of size 256×256. These images have been processed..."

L 260 *'We have demonstrated in Appendix C' Using the past tense feels a bit weird here. I would recommend using the present.*

- Done.

L 267 'to check the Courant-Friedrichs-Lewy (CFL) condition' → 'to satistfy the Courant-Friedrichs-Lewy (CFL) condition'

- Done.

L 270 *'It takes previous observations'. How many observations in the past? How do you merge the information from all these observations? Are they stacked in the channel direction?*

- We use the last 4 images, they are stacked in the channel direction. We added this information in the text:

"It takes ~~previous observations~~ the last four observations stacked on the channel axis, and estimates ..."

L 273 *'doesn't' → 'does not'.*

- Corrected here and in two other places.

L 316-317 *'with a total of approximately 100,000 images.' I think that it would be better and in that case even shorter to give the exact number.*

- The exact number is 105,120, we added this information in the text:

"The training was carried out on a dataset containing about three years of data from 2017 to 2019, with a total of ~~approximately 100,000~~ 105 120 images."

L 320 *'After cleaning' Please describe this cleaning step. Furthermore, please also describe what rule you use to split between training and validation.*

- The cleaning was done by removing the images with zero cloud cover and gathered all the sequences with 12 constructive images. The split between training and validation was done randomly. We added this information in the text and moved the sentence "To improve ... patterns." to the end of the paragraph, and divided it into two sentences:

"The training was carried out on a dataset containing about three years of data from 2017 to 2019, with a total of ~~approximately 100, 000 images.~~ 105 120 images. The images with zero cloud cover were removed, then we assembled all the sequences with 12 consecutive images. After this cleaning step, we randomly split the dataset, 8 224 sequences were used for training, and 432 for validation. The test set was performed on a separate dataset from the same region but from 2021.

To improve the diversity of the training set and take into account a possible overfitting on the typical movements of clouds in the Western Europe region, we randomly applied simple transformations to the images, more precisely, rotations of 90, 180 and 270 degrees, which increased the dataset size and improved the model's ability to learn various cloud motion patterns.~~After cleaning, about 8000 sequences of 12 images were used for training and about 400 sequences for validation. The test set was done on a separate dataset from the same region but from the year 2021.~~"

L 404-405 *'In the comparative evaluation, we included the widely used U-Net' If I understood correctly, you used as baseline a 'vanilla' U-Net. Why not using a U-Net Xception style as for HyPhAI?*

- We used the classical U-Net as a baseline because it is the one that is used is other works for the same task. We added this information in the text:

"In the comparative evaluation, we included the ~~widely used~~ well-known U-Net ~~(~~[Ronneberger et al., 2015]. This classical U-Net is different from the one used to estimate the velocity in the proposed hybrid models (refer to Fig. 3 and Fig. 4). The choice of this classical U-Net for comparison is justified by the fact that it is the most widely used in the literature for the same task (e.g. [Ayzel et al., 2020, Berthomier et al., 2020, Trebing et al., 2021]~~, e.g.~~)."

Figure 7 *While I agree that confidence intervals are in general needed, here they make the figure unreadable: there is just too much information. Furthermore, what confidence intervals are these:* 99%?95%?90% *? other?*

- We believe that the confidence intervals are important to show the uncertainty in the scores. However, we understand that the figure is too crowded, thus we removed them and provided another figure with the confidence intervals in the Appendix. The treshold used for the confidence intervals is 99%. We added this information in the caption:

"The confidence intervals were estimated using Bootstrapping, with a threshold of 99%."

Table 1 *Please describe what bold font means.*

- The bold font indicates the best score. We added this information in the caption:

"... (↑: higher is better, ↓: lower is better). The best scores are indicated in bold font.

Figure 8 *Labels are too small on this figure. What class correspond to colour 'beige' (which can be seen e.g. North-East of France)? Finally, the projection for this map seems a bit weird (possibly flattened in the latitude direction).*

- We increased the size of the labels. The beige colour is not in the labels, it corresponds to the land areas. This information is added to the figure's caption. The projection used here is the plate carrée, hence the effect noticed in the reviewer's comment.

L 440-441 *'highlighting that HyPhAI-1 produces more realistic and less blurry forecasts compared to the U-Net' Rigorously speaking this statement is true, but in my opinion it is a bit misleading because it hides the fact that even with HyPhAI-1 the prediction are much smoother than the truth.*

- We agree that point that the HyPhAICCast-1 predictions are smoother than the ground truth, and this is admitted in the conclusion. But here we are comparing the HyPhAICCast-1 predictions with the U-Net predictions, and we explained in the same paragraph the reason behind the HyPhAICCast-1 loss of details:

"~~, the~~ The lost details in ~~HyPhAI-1~~HyPhAICCast-1's predictions are only due to the ~~diffusion added numerically by the discretisation scheme used~~ numerical scheme, in ideal conditions, the HyPhAICCast-1 should preserve the same details during the advection process, and there is no other trainable part in between that can smooth the predictions; however, the upwind discretisation used scheme adds a numerical diffusion and crushing the small cloud cells (refer to ~~D~~ Appendix E for more details). "

L 445-446 *'the lost details in HyPhAI-1's predictions are only due to the diffusion added numerically by the discretisation scheme used' Are you sure about this statement? As far as I know, many ML models trained with the point-wise metrics tend to yield smooth predictions because of the double penalty issue (see, e.g., Bonavita 2023). I suspect that this is the case in your model. If not, and hence if numerical diffusion is the only obstacle, then can't you use another numerical scheme with less numerical diffusion?*

- We understand the suspicion of the reviewer. However, the only trainable component of HyPhAICCast-1 is the one used to estimate the velocity field, and this component does not have direct access to the loss function. The velocity field is used to advect the cloud cover field. In ideal conditions, we should preserve the same details during the advection process, and there is no other trainable part in between that can smooth the predictions. This information is added to the same paragraph.

"~~, the~~ The lost details in ~~HyPhAI-1~~HyPhAICCast-1's predictions are only due to the ~~diffusion added numerically by the discretisation scheme used~~ numerical scheme, in ideal conditions, the HyPhAICCast-1 should preserve the same details during the advection process, and there is no other trainable part in between that can smooth the predictions; however, the upwind discretisation used scheme adds a numerical diffusion and crushing the small cloud cells (refer to ~~D~~ Appendix E for more details). "

- Regarding the use of another numerical scheme, it is worth noting that the scheme to use should be automatically differentiable, stable, and preserves both details and the probabilistic properties. At the moment, we do not have an alternative scheme that satisfies all these conditions. For example, the central differences preserve more details than the upwind scheme, even if we ignore the dispersion issues, but it is not stable for the last lead times as the dispersion issues become more important.

Figure 9 *Why did you use a different extent for this map?*

- As the figure shows one image, we have more space to show it with a larger extent, and the choice of the orthographic projection here is more aesthetic.

Figure 10 *I think that figure 10 is discussed after figure 11 in the text.*

- Corrected.

Section 4.4 *Can we really draw robust conclusions here by looking at the validation scores (and not the test scores)? Furthermore, the fact that hybrid models are usually more accurate with less training data is already widely known in the hybrid modelling literature (see, again, [Cheng et al., 2023].*

- We understand the concerns of the reviewer. However, validation data are generally used to tune hyperparameters, and we consider the required data size to be one of these hyperparameters. And we presented the results of these experiments, we believe that we can draw robust conclusions based on them as these data are not seen by the model during training. Regarding the second point, we agree that it is a well-known fact that hybrid models can be data-efficient [Schweidtmann et al., 2024, Cheng et al., 2023], and we are not claiming this. We clarified this point in the text:

"This finding indicates that this hybrid model is remarkably data efficient, capable of delivering satisfactory performance even with limited training data, which has been highlighted by other studies [Schweidtmann et al., 2024, Cheng et al., 2023]. This quality is very important, particularly for tasks with insufficient provided data."

- However, we believe that it is important to show that in our context, and we provided a quantitative mesure of this efficiency.

L 476 *'earth' → 'Earth'.*

- Corrected here and in one other place.

L 478 *'This expansive full disk domain is 14 times the size of the training area.' If I am not mistaken, the original domain is* 256×256 *and the full disk is* 3712×3712 *(BTW, this information is only mentioned in the caption of Fig. A3, it would be better to mention it in the text). The full disk is therefore 210.25 times bigger, right?*

- Yes, it is 210.25 times larger (14 times in each direction). We changed this information in the text. The size of the full disk is now mentioned in the text:

"~~This expansive full disk domain is 14 times~~ The satellite observations of this expansive full-disk domain are of size 3712×3712, which is 210.25 times larger than the size of the training ~~area~~ones."

Figure 12 *This figure is not referenced in the text.*

- We added a reference to this figure in the text:

Despite the significant differences between the training domain and the full disk, we observed ~~a remarkable adaptation of the HyPhAI-1 model to this new context~~ good qualitative forecasts of the HyPhAICCast-1 model on this new domain without any specific training on it (see Fig. ~~A3~~12 and Fig. A4).

References

[Ayzel et al., 2020] Ayzel, G., Scheffer, T., and Heistermann, M. (2020). RainNet v1.0: a convolutional neural network for radar-based precipitation nowcasting. *Geoscientific Model Development*, 13(6):2631–2644. Publisher: Copernicus GmbH.

[Berthomier et al., 2020] Berthomier, L., Pradel, B., and Perez, L. (2020). Cloud Cover Nowcasting with Deep Learning. In *2020 Tenth International Conference on Image Processing Theory, Tools and Applications (IPTA)*, pages 1–6. arXiv:2009.11577 [cs].

[Cheng et al., 2023] Cheng, S., Quilodrán-Casas, C., Ouala, S., Farchi, A., Liu, C., Tandeo, P., Fablet, R., Lucor, D., Iooss, B., Brajard, J., Xiao, D., Janjic, T., Ding, W., Guo, Y., Carrassi, A., Bocquet, M., and Arcucci, R. (2023). Machine Learning With Data Assimilation and Uncertainty Quantification for Dynamical Systems: A Review. *IEEE/CAA Journal of Automatica Sinica*, 10(6):1361–1387. Publisher: IEEE/CAA Journal of Automatica Sinica.

[Joe et al., 2022] Joe, P., Sun, J., Yussouf, N., Goodman, S., Riemer, M., Gouda, K. C., Golding, B., Rogers, R., Isaac, G., Wilson, J., Li, P. W. P., Wulfmeyer, V., Elmore, K., Onvlee, J., Chong, P., and Ladue, J. (2022). Predicting the Weather: A Partnership of Observation Scientists and Forecasters. In Golding, B., editor, *Towards the "Perfect" Weather Warning: Bridging Disciplinary Gaps through Partnership and Communication*, pages 201–254. Springer International Publishing, Cham.

[Matte et al., 2022] Matte, D., Christensen, J. H., Feddersen, H., Vedel, H., Nielsen, N. W., Pedersen, R. A., and Zeitzen, R. M. K. (2022). On the Potentials and Limitations of Attributing a Small-Scale Climate Event. *Geophysical Research Letters*, 49(16):e2022GL099481. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1029/2022GL099481.

[Pavliotis and Stuart, 2008] Pavliotis, G. and Stuart, A. (2008). *Multiscale Methods: Averaging and Homogenization*, volume 53. Springer, New York, NY.

[Ronneberger et al., 2015] Ronneberger, O., Fischer, P., and Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. In Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, volume 9351, pages 234–241. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

[Schultz et al., 2021] Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., and Stadtler, S. (2021). Can deep learning beat numerical weather prediction? *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 379(2194):20200097.

[Schweidtmann et al., 2024] Schweidtmann, A. M., Zhang, D., and von Stosch, M. (2024). A review and perspective on hybrid modeling methodologies. *Digital Chemical Engineering*, 10:100136.

[Trebing et al., 2021] Trebing, K., Stanczyk, T., and Mehrkanoon, S. (2021). SmaAt-UNet: Precipitation nowcasting using a small attention-UNet architecture. *Pattern Recognition Letters*, 145:178–186.