# HyPhAI v1.0: Hybrid Physics-AI architecture for cloud cover nowcasting – Review report

6th May 2024

In this manuscript, the authors propose a new method that combines a neural network and a numerical integration scheme for cloud cover nowcasting. The innovative part of the method is that it is built on top of a cloud cover classification and hence relies on probabilistic approaches (in particular the network is trained using cross-entropy). The method is applied to a large dataset of real satellite images which have been labelled beforehand.

Overall I found the manuscript well written and easily understandable, with interesting methodological developments. I am therefore really positive about this manuscript.

## 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.

### 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.

### 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.

### 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 $3712 \times 3712$ squared image?
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)?

## 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.

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

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

**L 52** 'the hybridisation available techniques' → 'the available hybridisation techniques'?

**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).

**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.

**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.

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

**L 122-123** 'The choice of $l$ depends, among other things, on the statistical model $f_\theta$.' 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.

**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.

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

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

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

**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?

**L 273** 'doesn't' → 'does not'.

**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.

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

**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?

**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?

**Table 1** Please describe what bold font means.

**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).

**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.

**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?

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

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

**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).

**L 476** 'earth' → 'Earth'.

**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 $256 \times 256$ and the full disk is $3712 \times 3712$ (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?

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

## References

Bonavita, Massimo (2023). 'On Some Limitations of Data-Driven Weather Forecasting Models'. Version 2. In: DOI: 10.48550/arxiv.2309.08473.

Cheng, Sibo et al. (June 2023). 'Machine Learning With Data Assimilation and Uncertainty Quantification for Dynamical Systems: A Review'. In: *IEEE/CAA Journal of Automatica Sinica* 10.6, pp. 1361–1387. DOI: 10.1109/JAS.2023.123537.