Towards physically consistent data-driven weather forecasting: Integrating data assimilation with equivariance-preserving spatial transformers in a case study with ERA5

The present manuscript introduces a new deep learning framework to forecast global geopotential height. More specifically, the authors introduce a U-NET (Sec 3.1, Fig 1) using circular convolutions (Sec 3.1.1) and augment it with an equivariance-preserving module (U-STN, Sec 3.1.2) to improve the overall accuracy of the forecast (Fig 3) and the consistency of the predicted patterns (Fig 4). They couple the resulting network with a sigma-point Ensemble Kalman filter (SPEnKF, Sec 3.2) that allows to assimilate "noisy observations" every 24 hours (Fig 5) or "virtual observations" every 12 hours produced by the same network run with a longer timestep (Fig 7). Evaluated using hourly, coarse-grained (Sec 2), ERA5 [2] meteorological reanalysis of 500-hPa geopotential height (Z500) from WeatherBench [3], the equivariance-preserving network using the SPEnKF to assimilate both "virtual" and "noisy observations" improves the performance of the same framework only assimilating "noisy observations" (Fig 8+9).

The manuscript is generally well-written, well-referenced, logically-structured; its figures are clear and its (surprisingly simple) code is accessible on GitHub (`https://github.com/ashesh6810/DDWP-DA`) and properly shared via Zenodo (`DOI10.5281/zenodo.4646676`). Given the methodological novelty and applicability of the U-STN+SPEnKF framework to data-driven weather forecasting, I recommend eventually publishing the present manuscript in *Geoscientific Model Development* (GMD). That being said, the article's impact may be hindered by incomplete benchmarking 1.1, a lack of testing on other meteorological variables 1.2, little justification of the equivariance-preserving module 1.3, and overly technical writing that may not be appropriate for GMD's audience 1.4. More details 1 and minor comments 2 are given below. I am optimistic that once improved, the manuscript will be a welcome addition to GMD and a helpful contribution to the sub-field of data-driven weather forecasting.

Contents

1 Major Issues

1.1 Benchmarking

[L220-244, Fig3] The manuscript's premise is that adding the equivariance-preserving module may improve the accuracy of data-driven weather forecasting, which is demonstrated by training a U-NET with and without the equivariance-preserving module and showing the resulting improvement in accuracy (as measured by the anomaly correlation coefficient) for lead times between 12 and 240 hours. This raises several issues:

- Despite using a well-defined benchmark (WeatherBench, [3]), the root mean squared error (RMSE) in is never calculated for the predictions without data assimilation (U-NET/U-STN), which prevents objective comparison with the baselines listed in Figure 2/Table 2 of [3]. I recommend at least calculating the RMSE in Z500 for lead times of 3 and 5 days to put the manuscript's results into the context of existing results (e.g., do U-NET/U-STN beat the simple linear regression leading to $\mathrm{RMSE}_{Z500}\,(3\mathrm{days}) \approx 693\mathrm{m}^2\mathrm{s}^{-2}$ **without data assimilation**? If the authors consider that only Z500 should be used as a predictor, then how do U-NET/U-STN perform compared to the linear regression equivalent that only uses Z500 as a predictor?).

- Once put into the WeatherBench context, it remains unclear whether U-STN systematically improves upon U-NET or if the result depends on the single set of (hyperparameters, weights, biases) explored in this manuscript. For instance, the only sensitivity explored in Figure 3 is that to initial conditions while the only sensitivity explored in Figure 6 is that to the timestep Δt I recommend more thoroughly testing the addition of the equivariance-preserving module across:
  - Different weights and biases for a fixed set of hyperparameters by retraining U-NET/U-STN with different weights initializations and callbacks
  - Different hyperparameters by changing the convolutional and dense layers characteristics (number, width, kernel size) within the U-NET/U-STN architectures
  - Different architectures altogether: Would an equivariance-preserving module help an artificial neural network (with or without bottlenecks), simple linear models, etc.?
  In summary, I recommend conducting sensitivity tests to determine whether the paper's key conclusions hold across architectures, hyperparameters, and different weights/biases.

1.2 Testing the Framework on Other Meteorological Variables

[L105-110]

- Given that the manuscript's conclusions should apply to data-driven weather forecasting in general, I recommend testing the framework on a few more meteorological variables, especially given how easy it is to download variables from WeatherBench and how short the manuscript's repository code is. Natural choices would be variables benchmarked in WeatherBench, i.e. 850-hPa temperature (T850), 2m temperature (T2M) and total precipitation (TP).

- At the very least, the authors should discuss how appropriate equivariance-preserving spatial transformers are for thermodynamic variables like T850, which (in contrast to dynamic variables like Z500) directly respond to the strong planetary gradient in solar insolation. I recommend adding at least T850 to clarify the generality of the results e.g. presented in Figure 3.

1.3 Justifying and Explaining Equivariance

- [L130-140] In the other reference cited by the authors to justify using the spatial transformer module [4], the invariance under spatiotemporal translation, uniform motion, rotation/reflection, and scaling is justified for the Navier-Stokes and heat equation. However, when it comes to atmospheric dynamics, strong asymmetries exist in the horizontal (including but not limited to the Coriolis parameter for dynamical quantities, the solar insolation for thermodynamical quantity, and the land mass for all quantities). Therefore, I recommend carefully justifying why it would be appropriate to use an equivariance-preserving module in the text of subsection 3.1.2.

- [L233-236] Similarly, it would be helpful to more clearly justify/explain why equivariance-preserving networks would improve the representation of wave-breaking events, which are not rotationally nor translationally invariant. I

recommend more rigorously justifying that claim by e.g. zooming in Figure 4, adding more variables and network configurations, and not relying on "As discussed before" when the word "breaking" was simply listed in L120.

1.4 Making the Manuscript more Accessible to GMD's Audience

1.4.1 Presentation of the Sigma-Point Ensemble Kalman Filter

[L163-219] The authors adapt the Sigma-Point Ensemble Kalman Filter (SPEnKF, [1]) to augment their data-driven weather prediction framework with data assimilation (DDWP+DA). I found this description hard to follow because it lacks context and justification; I recommend revising the text to address the following questions:

- Do the authors strictly follow the derivations/methods of [1] or are there some key modifications to couple it to the ML prediction framework?
- Are analysis and "observations" used interchangeably here? In the affirmative, I would recommend sticking to one or the other.
- According to [1], SPEnKF is particularly well-suited for non Gaussian background/observation errors (e.g. multiplicative noise). Why then assume that ϵ be Gaussian, which (if I understand the derivation correctly) leads to Gaussian observational errors as $\boldsymbol{H} = \boldsymbol{I}$?

Additionally:
[L194] I recommend explicitly stating that representing observational noise using a random Gaussian process is a big approximation.
[L197] "with a certain level of uncertainty": I recommend explicitly stating that the uncertainty will be ideally represented by varying σ_{obs}.
[L205] If $\boldsymbol{P_{ab}}$ is the cross-covariance matrix between the ensemble and observations, shouldn't the two last Z_{ens} be Z_{obs} in equation (8)?

1.4.2 Overly technical vocabulary used throughout the manuscript

GMD is targeted at the geosciences community: Although the community is relatively quite proficient in computational science, a lot of the vocabulary and technical terms used throughout the manuscript makes it difficult to read without ML background. To make the manuscript more accessible to the geoscientific community, I recommend:

- Quickly defining technical ML/DA terms used throughout the manuscript the first time they are introduced. This includes but is not limited to: convolutional neural network, deep spatial transformer, equivariance, Ensemble Kalman filter, encoding/decoding, autoregressive models, etc.
- Alternatively, adding a "ML definition" Table to the manuscript.
- Using more intuitive acronyms. For instance, U-STN, SPEnKF, and DDWP+DA are not particularly intuitive and may force the readers to go back and forth when reading the manuscript.

Also see comments in 2 to improve the manuscript's accessibility.

1.5 Reproducibility

1.5.1 Unet's architecture

[L113-L122] After checking the U-NET's architecture at `www.github.com/ashesh6810/DDWP-DA/blob/master/Unet_STN.ipynb` (same script as the one shared via Zenodo if I am not mistaken), I noticed that Figure 1 was not representative of the architectures used for U-NET/U-STN, which include additional dense layers after the convolutional layers. Additionally, because the authors do not disclose the type of pooling layers used in Figure 1, the architecture of the main algorithms used in the manuscript cannot be reproduced from the text.

- As the authors cite [5], I recommend following their Table 1 to transparently share the U-NET's architecture.
- Additionally, it would be nice to explicitly list the differences between [5] and this manuscript's U-NET, including (but not limited to) the presence of dense layers) to facilitate the comparison between the two frameworks.

1.5.2 Weights and Biases

[L113-122] The weights and biases of the neural networks are not shared (to my knowledge) in the code's repository, making the manuscript non reproducible. I highly encourage the authors to share the weights and biases of their networks for reproducibility purposes.

1.5.3 Equivariance-preserving Module

[L131-140] The authors do not provide enough details to help readers implement the spatial transformer module. I recommend explicitly stating how to implement this module (the Bilinear Interpolation Class at `https://github.com/ashesh6810/DDWP-DA/blob/master/layers.py`). This could be done by e.g.:

- adding an "algorithm" in the manuscript's text, and
- giving more context for why the spatial transformer module requires adding a bi-linear interpolation kernel between the convolutions and the up-sampling.

2 Minor Comments

[L37-39] Although it has been demonstrated for low-dimensional systems in fluid dynamics, it is not trivial that:

- Incorporating physical constraints into a physics-agnostic data-driven weather prediction framework would require less data and hence remedy the short training set problem,
- the equivariance-preserving spatial transformer introduced in this manuscript can be used to enforce physical constraints.

I recommend rephrasing these introductory sentences or carefully justifying these two claims.
[L97] Is "data-drivenly" correct?
[L94-99] These three points, especially the third one, are extremely technical and hard to understand without re-reading them several times. Would it be possible to rephrase them?
[L105-110] This section is extremely short: Would it be possible to

- add more context for why the authors first decided to test the framework on Z500 specifically,
- add the number of samples for each training set, and
- add a short justification for the training/validation/test split chosen by the authors?

[L154] becomes a U-NET → becomes a standard U-NET?
[L160] (Over the baseline, U-NET) → Benchmarking against another quick fit by the authors is far from rigorous: Following the major comment 1.1, would it be possible to add a subsection to Section 2 or at least a paragraph in Section 3.3 to describe and justify the paper's benchmarking methods?
[L164] "unscented transformation" requires more context for readers who are not versed in the Ensemble Kalman filter
[L177-178] ~50-100 members are used → Missing reference: Are the authors referring to the Integrated Forecasting System?
[Fig2 caption] "DA ... DDWP" → Consider spelling out acronyms or rephrasing to facilitate the caption's readability.
[L193] $k \in [-D, -D+1, ...D-1, D]$ → Do the authors mean $k \in \{-D, -D+1, ...D-1, D\}$ or equivalently $k \in [\![-D, D]\!]$?
[L262-264] I find this claim confusing:

- Is ERA5 truly noise-free?
- Doesn't the de-noising property come from the fact that U-STN is a deterministic neural network, which by definition cannot produce noise?
- Or are the authors referring to the fact that U-STN has a filtering effect that makes the normalized output variance smaller than the normalized input variance? If that is the case, I recommend clarifying the text and quantitatively justifying this claim about U-STN.

[L273-275] This qualitative explanation ignores the fact that errors made by the neural network are larger (in physical units) for larger Δt. Therefore, I recommend clarifying that the error accumulation is larger than the error increase with Δt. Would it be possible to quantitatively verify that claim (e.g. via a supplemental figure)?
[L337] "variational DA algorithms": Which variational DA algorithms are the authors referring to? Ideally, provide references for readers who are less familiar with DA.
[L347-348] Would it be possible to add the GitHub repository's link as it may be more convenient than downloading the archived code for some readers?

References

[1] J. T. Ambadan and Y. Tang. Sigma-point particle filter for parameter estimation in a multiplicative noise environment. *Journal of Advances in Modeling Earth Systems*, 3(4), apr 2011.

[2] H. Hersbach and H. The ERA5 Atmospheric Reanalysis. *American Geophysical Union, Fall General Assembly 2016, abstract id. NG33D-01*, 2016.

[3] S. Rasp, P. D. Dueben, S. Scher, J. A. Weyn, S. Mouatadid, and N. Thuerey. WeatherBench: A benchmark dataset for data-driven weather forecasting. feb 2020.

[4] R. Wang, R. Walters, and R. Yu. Incorporating Symmetry into Deep Dynamics Models for Improved Generalization. 2020.

[5] J. A. Weyn, D. R. Durran, and R. Caruana. Improving Data-Driven Global Weather Prediction Using Deep Convolutional Neural Networks on a Cubed Sphere. *Journal of Advances in Modeling Earth Systems*, 12(9):e2020MS002109, sep 2020.