Geoscientific Model Development Discussions

Geosci. Model Dev. Discuss., referee comment RC2
https://doi.org/10.5194/gmd-2021-430-RC2, 2022

Comment on gmd-2021-430

Anonymous Referee #2

Referee comment on "Temperature forecasting by deep-learning methods" by Bing Gong et al., Geosci. Model Dev. Discuss., https://doi.org/10.5194/gmd-2021-430-RC2, 2022

This is a review of the paper "Temperature forecasting by deep learning methods" by Bing Gong, Michael Langguth, et al.

This paper describes the use of an existing generative adversarial neural network architecture and approach for video prediction, SAVP, to the problem of predicting the evolution of the temperature field over central Europe. While the results are not yet competitive with operational weather forecasts, the input data is relatively coarse and very few predictors are used, so this is not surprising. The authors perform several ablation studies to identify the main contributions to the model's strength, although I have some minor criticisms about the details of some of these. Nevertheless, I believe the paper represents an interesting extension to the existing literature, within the area of purely data-driven approaches to weather forecasting.

I have three main comments about the authors' approach:

1) The authors use a generative model, with an explicit sampling step, which allows them to generate multiple forecasts for a single input. However, the authors do not seem to explore this aspect at all, apart from a brief mention of probabilistic prediction in the conclusion section. There are a large number of ensemble verification metrics avilable to assess the calibration of the generated ensemble, i.e., to see to what extent the ground truth is interchangeable with a generated ensemble member. Some simpler ones include spread-skill plots and ensemble rank histograms. This be a route that the authors do not wish to pursue yet and leave for future work, but it might be interesting to at least have some idea of how different the generated sequences can be for the same given input data, even for one or two case studies.

2) Regarding predictors, am I right in thinking that the network is given no direct information indicating where in the diurnal cycle it is starting to forecast from? E.g. time of day and day of year, or total incoming solar radiation, etc.? I.e. it has to infer this from

the patterns seen in the first 12 hours' data? If so, this seems like a strange choice, and one might imagine the model occasionally becoming confused by unusual temperature variations in the first 12 hours. Are there any signs that something like this happens? More generally, if you look at some of the worst predictions (e.g. by average MSE over the 12 hours), is there anything interesting about the failure modes, which may hint at extra predictors to use? I imagine the authors may wish to use a much larger set of meteorological variables in future work!

3) Regarding the experiment that varied the domain size, I understand the authors believe that the varying domain (which the metrics are computed over) contributes majorly to the difference in scores -- the larger domains have larger proportions of water, which leads to lower MSE, etc. As a result, I don't feel this part of the paper contributes much insight in its current form. Can I suggest that the evaluation is performed on the same physical domain each time, e.g. the 72 x 44 central region? I.e., when the larger domains are being used, they are cropped to the central 72 x 44 region before various metrics are calculated. In this way, the comparison is fairer, and the effect of 'larger context' can be isolated from the varying evaluation domain.

For similar reasons, I am somewhat skeptical of the 'sensitivity to number of years of training data' result, since (if I understand correctly) the evaluation is performed on three different years. These themselves may be more or less difficult to predict. If it is feasible to re-run this part of the work to avoid evaluating on different years, this would seem like a good idea. If not, I suggest they at least add a corresponding caveat to the results discussion!

Minor comments:

1) What is the ConvLSTM model trained on? I couldn't spot this easily in the text. Is it just trained to minimise MSE (i.e., L^2 error)?

2) I believe the original ConvLSTM paper is normally cited as Shi et al. (2015), not Xingjian et al. (2015)?

3) In Figure 5 (and similar figures), I assume the three lines for each model correspond to the three different datasets (evaluation/training years) used? This could be made a bit clearer, e.g. in the caption.

Finally, here are a few small typos/grammatical mistakes, etc., that I spotted:

Line 17: as additional predictor -> as an additional predictor

Line 206: of a 24 time steps -> of 24 time steps
Line 207: This results into about -> This results in about
Line 235: which encodes -> which encode
Line 238: no comma needed after 'both'
Line 257: condinoned -> conditioned
Line 258: missing Z after 'latent space'
Line 467: for a 12-hour forecasts, is attained -> for a 12-hour forecast is attained
Line 468: higher spatial solutions -> higher spatial resolution
Line 480: repeated word 'motivate'
Line 485: deep neural can -> deep neural networks can
Line 496: into -> in
Line 518: as list in -> as listed in
Line 521: ration -> ratio
Line 525: I + J -> I x J
Line 533: and each of the day -> and each hour of the day
Line 560: I think 'disposal' should be something else, but I am not sure what?