General statement

We would like to thank the editor for coordinating the review of our work. We would also like to thank the two reviewers for their helpful and valuable comments on our study. In the following, we point to point addressed the referees' comments and revised the manuscript accordingly. For clarity, our responses are highlighted in red.

Referee comment #1

The present paper presents the use of a stochastic adversarial video prediction model to forecasting the two-meter temperature. While the paper is interesting to read and the conclusions seem valid, I do have several points that I think would have to be addressed before the paper could be considered for publication in GMD. In particular:

1) The present paper is a contribution to an increasing line of research on the application of a deep-learning based methodology to weather prediction, in particular the use of an existing video prediction model applied to weather prediction. This line of work has been of great interest when first showcased through the contributions of Weyn et al. and Dueben et al., just to name a few, but it does not feel that the present paper adds anything substantially new besides using a different architecture for the same problem. One main issue is that meteorological data is fundamentally different from generic video data in that it follows a well-defined system of partial differential equations. In this purely data-driven approach it seems that one has to be willing to throw away more than a hundred years of research on the understanding of these governing equations of hydro-thermodynamics just to be able to use an off-the-shelf video prediction architecture, which does not seem to come close to where traditional numerical methods can go today in terms of relevant forecast metrics. The question to ask is hence whether it is indeed the right approach going forward, or whether one should strive to combine data-driven approaches with the inductive bias as provided by the fundamental laws of physics. There is a growing interest in physics-informed machine learning, which allows combining differential equations with data-driven machine learning which in a way seems more appropriate for the present problem at hand. If that was possible for the present model then I think the paper would become much stronger and more suitable for what would actually be required for weather prediction.

Many thanks for sharing your thoughts on the overall method of data-driven weather forecasts with the help of video prediction methods from computer vision.
We agree that physical informed neural networks (PINNs) constitute an appealing approach, especially for any physical process such as the evolution of the atmospheric state. While appealing due to the explicit awareness of physical laws, PINNs also have their drawbacks, especially for systems with a high-frequency and multi-scale nature (see, e.g., Wang et al., 2020, Fuks and Tchelepi, 2020, or Jin et al., 2021). The evolution of the atmosphere can be described by a set of partial differential equations (PDE) that constitute the Navier Stokes equation, the 1st law of thermodynamics and the continuity equation of a multi-phase fluid. This PDE is inherently multi-scale and involves various high-frequency processes (e.g. gravity waves) which make the application of PINNs extremely challenging. As a consequence, mainly strongly simplified versions of this set of PDEs have been tackled with PINNs (see, e.g. Raissi et al., 2019 and Rao et al., 2020).

For our particular case of the 2m temperature, the formulation of simplified PDEs or alternatively of physical constraints is far from straightforward, since the 2m temperature is subject to various physical processes down to millimetre scale, e.g. surface heat fluxes and turbulence in the (near-) surface layer. Note that the 2m temperature is usually also not a prognostic variable in atmospheric models. It rather constitutes an empirical fit between the temperature in the first model layer (located about 10m above ground) and the skin temperature, see e.g. the Forecast User Guide by ECMWF.
Due to these arguments, we argue that data-driven methods are still a valid approach for the application of neural networks and as detailed below, we also believe that our study provides additional value with respect to previous studies such as Dueben et al., 2018, Rasp et al., 2020, and Weyn et al., 2020. Nevertheless, we emphasize the above statement in our revised manuscript

2) The selection of features (cloud cover, 850 hPa temperature and two-meter temperature) seems slightly arbitrary. While the authors do provide some justification for the selection of these parameters, there are many more parameters that influence the evolution of the two-meter temperature. The authors then state "A more systematic variable selection process as is typical for data science studies is beyond the scope of this paper.", but I do not believe this is a justifiable statement here, because the authors do carry out a data science study in this paper. Again, if this was the first paper to be written on using a video prediction model for weather prediction this point could be easily forgiven, but as there are many other papers out there that provide proof of concept that such models can predict the future weather to some degree I think some more work needs to be done here to justify this feature selection, and to show which features have to be selected to get the best possible model results. In the machine learning literature it is customary to carry out ablation studies that showcase the importance of various aspects of the data/model components being used, and I think such a study would be beneficial here as well.

We acknowledge the lack of justification for our feature selection. For the revised manuscript, we have provided more reasoning including some proper references (See line 192-200 in revised manuscript).

Additionally, we would like to highlight the added value, that is the inclusion of additional predictors based on expert knowledge. Previous studies mainly only used the 2m temperature itself as a predictor for their task (e.g. Bihlo, 2020, Rasp et al., 2020). Due to this, the neural network is enforced to encode the relevant atmospheric state from one state variable at a single height above the surface. By including the temperature at 850 hPa (T850) and the total cloud cover (TCC), we inform the network on the temperature of the air mass (850 hPa temperature is not directly affected by the interaction with the surface) and include a key driver of the surface heat fluxes (modulation of incoming short-wave and outgoing long-wave radiation).

In addition, we would like to mention that our study did several ablation studies to investigate the relevance of different data/model components. For instance, we control the scaling factor of reconstruction loss, and to figure out how does the model perform when increasing weights on GAN component. The results were demonstrated in the Section 4.2. We also carried out several sensitivity analyses in Section 4.3 to data components.

One of our sensitivity studies reveals that including T850 is beneficial, while TCC only yields minor improvements. Replacing TCC for the sake of a more informative predictor is of course an option, but it would be computationally expensive due to re-running the training step.

While we agree that more informative predictors can be beneficial to forecasting and a systematic search could be warranted. Nevertheless, the setting of our current experiment has arrived at the computation limitations (12 hours of inputs and 12 hours of output). For the sake of releasing the memory constrains, we investigated whether reducing the input hours could preserve the model accuracy in our revised manuscript (see Line 443 - 450). The results demonstrate it is necessary to give the half-diurnal cycle of 2m temperature as inputs.

To make more variables as inputs feasible, we are now investigating the use of different network architectures which are more memory-efficient – such as Tansformer-based networks. These would allow a light-weight encoding of sequence data compared to the usage of recurrent layers (Vaswani A, etc 2017). However, this clearly has to be seen as a new study and goes beyond the scope of the present paper.

3) The baseline comparison model used is a standard convolutional LSTM model. This is the simplest possible model for video frame prediction and it is well-known to perform rather poorly as it exhibits an excessive amount of diffusion. Thus, beating this baseline is rather straightforward so I wonder if the comparison of the authors' model to this simple model really yields a lot of information about the absolute strength of this model. It would be great to add a somewhat more state-of-the-art comparison model as well to be truly able to assess how good the stochastic adversarial video prediction model is for the present problem. Related to this, it would also be useful to add the performance metrics of traditional weather forecasting models as point of comparsion. Right now this information is just provided in the Discussions section but it would be nice to show these metrics in the plots as well.

Thank you for this suggestion. In our updated vision, we added the ERA5 short range forecast as a reference model for comparing to our models and we have a further discussion on the results (see Figure 5 in the revised manuscript).

4) Owing to the interest in data-driven weather forecasting, a standard benchmark "WeatherBench" has been proposed to facilitate comparison with other deep learning-based models. The present paper does not use this benchmark but rather investigates the model performance over Europe instead. This makes positioning this work within the wider literature rather challenging so I wonder if it would not be better to provide these results instead (or in addition) for the WeatherBench dataset as well. Again, this would facilitate comparison with other approaches that have been proposed for data-driven weather forecasting.

Regarding the WeatherBench, we would like to emphasize that WeatherBench mainly focuses on medium-range forecasting with a long time step of 6 hours in the iterative forecasts approach (see Fig. 1 in Rasp et al., 2020). Our study focuses on short-range forecasts with a small-time step of one hour which matches the output time step of NWP models. This makes the task more challenging for autoregressive models since errors accumulate (see Table 2 in Rasp et al., 2020 which reveals much higher RMSE for the iterative forecasts models).

Due to the different issues we are tackling, we refrain from integrating and testing our ML models on WeatherBench dataset in our updated manuscript. Alternatively, we tested the CNN model architecture described as a baseline model in (Rasp et al. 2020) (see Line 260-265 line in revised manuscript) and compare the results against our tested model architectures, the simple ConvLSTM and the SAVP model.

Concretely, we applied the best-performing CNN on the WeeatherDataset in an iterative forecasting mode.The mean square error as loss function to is applied to optimize three input variables of one preceding hour (2m temperature, total cloud cover, and temperature at 850 hPa), The CNN model is compared to the persistence forecasts as demonstrated in the Figure (A1). The iterative CNN forecasts can beat persistence forecasts up to 4 leading hours. However, after 4 hours, the model errors accumulate fast and diverge quickly. It completely deteriorates and the errors explodes after lead time of 10 hours.

This clearly reveals that the CNN model fails to beat our baseline model-convLSTM, which indicates 2D convolutional neural networks fail to capture the temporal dependency and to obtain skillful forecast for longer lead times. We updated this results and discussion accordingly in our revised manuscript (Line 365 - 372)

In summary, while the present paper is interesting to read I do believe there isn't a sufficient amount of novelty yet that warrants publication in GMD in its present form. The main contribution of picking a video prediction model and applying it to weather forecasting has been done several times in the recent literature so this does not feel novel enough anymore unless other open aspects of data-driven weather prediction are investigated in addition. These could be, as indicated above, a combination with differential-equations based models, a more thorough investigation of which parameters are responsible for the success of the proposed model, beating other existing approaches for the exact same problem domain, just to name a few.

We hope that our arguments listed above together with the revised revision of our manuscript provide the necessary degree of novelty and improvement on our study.

Referee comment #2

- 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.

Firstly, thank you so much for your time for reviewing our manuscript and for the comments to help us improve our work.

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.

Thank you so much for your suggestion. We fully agree that the SAVP model can be used for stochastic forecasting. The SAVP model can also be used for probabilistic forecasting through both noise samples as input and latent space of VAE component. Particularly, the encoder maps the input dimension to latent representations that follow normal distributions instead of single point. The decoder unfolds the latent code that samples from normal distributions, in which way to incorporate the stochasticity into the model. We would like leave this in the future study and limit our research to deterministic forecasting. In our updated manuscript, we further elaborate the probabilistic functions of SAVP model and how it can be used for future probabilistic forecasting (see line 515- 520) .

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!

Yes, we do not explicitly embed the daytime and the year of the date to the model. This means that the model has to infer the diurnal cycle in a purely data-driven way from the input sequence. This is also the reason why the input sequence comprises 12 hours so that at least one half of the diurnal cycle is inputted to the model. The task is then to construct the

second half based on the first half. A high-level example would be an input sequence from 4 to 15 UTC where the model would have to encode the daily warming until the afternoon which will be followed by cooling over the evening and in the night. Figure 6 provides an example from our test data. Yet, we have not seen special failure modes for the worst predictions.

However, in our revised manuscript, we conducted some experiments to further access this assumption by varying the input sequences (hours). We notice that given the half diurnal cycle of data (12 preceding hours), the model obtains the best forecasting accuracy in terms of MSE (see Figure 11). The performance slightly deteriorates with a lead time of 12 hours by reducing input hours from 12 to 5. The performance significantly drops from 5 to 2 hours. Similar results also hold for other evaluation metrics ACC and SSIM except for gradient ratio.

We also consider to include the daytime, season as adding inputs to be embedded into our models. However, adding this conditioning information to the input is conceptually not straightforward for a vanilla video prediction model such as SAVP. Since our intention was to check the performance of these models into another application, in our case weather forecasting, we refrained from revising the architecture on the input side and rather focussed on testing data-driven approaches. In addition, we note that a light-weight integration of this information would be required in our case since the current implementation of the SAVP model with three feature channels and a total sequence length of 24 hours occupies nearly all GPU memory.

The memory issue also motivates us to use a more memory-efficient model concept in the furture, namely a Transformer-based network which would allow us to embed or use more predictors in our follow-up studies (Vaswani A, etc 2017).

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.

Again thank you so much for your suggestion. For fair comparison, we re-evaluated our results on the same regions for test data in our revised manuscript accordingly.

For similar reasons, I am somewhat skeptical of the 'sensitivity to a 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!

Thank you for pointing this out. Indeed, for the sensitivity to the number of years of training data, we fixed the validation to 2016 and test the dataset to the year 2019 when evaluating the models with variations in the number of training samples. We had made this clear in the revised manuscript. (see line 405)

Minor comments:

1) What is the ConvLSTM model trained on? I couldn't spot this easily in the text. Is it just trained to minimize 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?

Thank you so much for providing this detailed list of typos and improper wording. We have revised our re-submitted manuscript accordingly.

Bibliography

Bihlo, Alex. "A generative adversarial network approach to (ensemble) weather prediction." *Neural Networks* 139 (2021): 1-16.

Dueben, Peter D., and Peter Bauer. "Challenges and design choices for global weather and climate models based on machine learning." *Geoscientific Model Development* 11.10 (2018): 3999-4009.

Fuks, Olga, and Hamdi A. Tchelepi. "Limitations of physics informed machine learning for nonlinear two-phase transport in porous media." *Journal of Machine Learning for Modeling and Computing* 1.1 (2020).

Jin, Xiaowei, et al. "NSFnets (Navier-Stokes flow nets): Physics-informed neural networks for the incompressible Navier-Stokes equations." *Journal of Computational Physics* 426 (2021): 109951.

Orlanski, Isidoro. "A rational subdivision of scales for atmospheric processes." *Bulletin of the American Meteorological Society* (1975): 527-530.

Raissi, Maziar, Paris Perdikaris, and George E. Karniadakis. "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations." *Journal of Computational physics* 378 (2019): 686-707.

Rao, Chengping, Hao Sun, and Yang Liu. "Physics-informed deep learning for incompressible laminar flows." *Theoretical and Applied Mechanics Letters* 10.3 (2020): 207-212.

Rasp, Stephan, et al. "WeatherBench: a benchmark data set for data-driven weather forecasting." *Journal of Advances in Modeling Earth Systems* 12.11 (2020): e2020MS002203.

Wang, Sifan, Xinling Yu, and Paris Perdikaris. "When and why PINNs fail to train: A neural tangent kernel perspective." *Journal of Computational Physics* 449 (2022): 110768.

Weyn, Jonathan A., Dale R. Durran, and Rich 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 (2020): e2020MS002109.

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. Advances in neural information processing systems. 2017;30.