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 will address the referees' comments and we will provide our plans and ideas for revising our manuscript. 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. Yet, 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.

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 plan to provide more reasoning including some proper references. 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).
Our sensitivity study 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. However, we believe that more informative predictors can be beneficial to forecasting. Nevertheless, the setting of our current experiment has arrived at the computation limitations (12 hours of inputs and 12 hours of output). To make more variables as inputs feasible, we will first reduce the input sequences and test the performance of the same lead time in the revised paper (see also reply to 2) of referee comment #2)..

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.

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.

We will certainly include the WeatherBench models as other baseline models. Additionally, we will compute the performance of the IFS forecasts as a reference in the evaluation section of our revised paper. However, we would like to emphasize that WeatherBench mainly focuses on medium-range forecasting with a long time step of 6 hours in the iterative 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 the models since errors accumulate (see Table 2 in Rasp et al., 2020 which reveals much higher RMSE for the ‘iterative’ models). Besides, our video prediction models have to learn explicitly the diurnal cycle of 2m temperature which constitutes the main temporal variability in our test study. This is for instance different from the study of Weyn et al., 2020, which also includes more predictors, but also uses a long time step of 6 hours.
Reducing the total lead time also allows us to reduce the size of the domain of interest since the variability is due to processes on local scale (correlation between time and spatial scale, see Orlanski, 1975). This in turn allows us to keep a much higher spatial resolution of 0.3° (compared to 5.625° in WeatherBench or 0.5° in Bihlo, 2021).

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 planned revision of our manuscript provide the necessary degree of novelty and improvement on our study.

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.