Temperature forecasting by deep learning methods

Bing Gong[1], Michael Langguth[1], Yan Ji[1], Amirpasha Mozaffari[1], Scarlet Stadtler[1], Karim Mache[1], and Martin G. Schultz[1]

[1]Jülich Supercomputing Centre, Forschungszentrum Jülich, 52425 Jülich, Germany

Correspondence: b.gong@fz-juelich.de

Abstract. Numerical weather prediction (NWP) models solve a system of partial differential equations based on physical laws to forecast the future state of the atmosphere. These models are deployed operationally, but they are computationally very expensive. Recently, the potential of deep neural networks to generate bespoke weather forecasts has been explored in a couple of scientific studies inspired by the success of video frame prediction models in computer vision. In this study, a simple recurrent neural network with convolutional filters, called ConvLSTM, and an advanced generative network, the Stochastic Adversarial Video Prediction (SAVP) model, are applied to create hourly forecasts of the $2\,\mathrm{m}$ temperature for the next 12 hours over Europe. We make use of 13 years of data from the ERA5 reanalysis, of which 11 years are utilized for training and one year each is used for validating and testing. We choose the $2\,\mathrm{m}$ temperature, total cloud cover and the $850\,\mathrm{hPa}$ temperature as predictors and show that both models attain predictive skill by outperforming persistence forecasts. SAVP is superior to ConvLSTM in terms of several evaluation metrics, confirming previous results from computer vision that larger, more complex networks are better suited to learn complex features and to generate better predictions. The 12-hour forecasts of SAVP attain a mean squared error (MSE) of about $2.3\,\mathrm{K}^2$, an anomaly correlation coefficient (ACC) larger than 0.85, a Structural Similarity Index (SSIM) of around 0.72 and a gradient ratio (r_G) of about 0.82. The ConvLSTM yields a higher MSE ($3.6\,\mathrm{K}^2$), a smaller ACC (0.80) and SSIM (0.65), but a slightly larger r_G (0.84). The superior performance of SAVP in terms of MSE, ACC and SSIM can be largely attributed to the generator. A sensitivity study shows that a larger weight of the GAN component in the SAVP loss leads to even better preservation of spatial variability at the cost of a somewhat increased MSE ($2.5\,\mathrm{K}^2$). Including the $850\,\mathrm{hPa}$ temperature as an additional predictor enhances the forecast quality and the model also benefits from a larger spatial domain. By contrast, adding the total cloud cover as predictor or reducing the amount of training data to eight years has only small effects. Although the temperature forecasts obtained in this way are still less powerful than contemporary NWP models, this study demonstrates that sophisticated deep neural networks may achieve considerable forecast quality beyond the nowcasting range in a purely data-driven way.

1 Introduction

Accurate predictions of weather are important for many aspects of modern society. They are of high relevance in economy and industry, e.g., for agriculture, for the (renewable) electric power industry or for prevention against natural hazards. Since the early 1960s, numerical weather prediction (NWP) models are run operationally at meteorological centers all over the world. These models are nowadays capable to simulate the dynamics of the global atmosphere down to the kilometre scale (Bauer

et al., 2015). While their predictions have reached a remarkable degree of reliability, the required computational resources are enormous (Zaengl et al., 2015).

Over recent years, deep learning (DL) has been successfully applied in computer vision applications, such as self-driving cars (Rao and Frtunikj, 2018), human action prediction (Kong and Fu, 2018), and anomaly detection (Liu et al., 2018), etc. These show that deep neural networks have the ability to recognize complex patterns and uncover highly non-linear relations in a data-driven way. Thus, hopes are raised that deep learning can be used for weather prediction and Earth system science (Schultz et al., 2021) which have to deal with many complex, multi-scale and non-linear coupled processes (Orlanski, 1975). The weather and climate communities are beginning to investigate the use of these advanced machine learning methods in the context of weather (McGovern et al., 2017) and climate forecasting (Reichstein et al., 2019), such as data assimilation (e.g. Hatfield et al., 2021), emulation of physical parameterization (e.g. Han et al., 2020), and detection of extreme weather events in climate datasets (e.g. Racah et al., 2016).

As discussed in Schultz et al. (2021), there are many potential applications of DL in the field of weather forecasting. DL methods can be integrated in each step of the NWP workflow which comprises pre-processing of observational data, assimilation of these data into the modelled real atmospheric state, forecasting with a numerical model and post-processing on the raw model outputs (see Fig. 1 in Schultz et al., 2021). Here, we provide a proof-of-concept study on replacing the NWP model with data-driven video prediction methods to forecast the evolution of the atmospheric state, particularly the 2 m temperature, up to 12 hours ahead. This is considerably longer than the typical range of nowcasting applications with a lead time of three hours or less (Wilson et al., 2010), but shorter than medium-range forecasts targeted in other studies (c.f. Scher, 2018; Rasp and Lerch, 2018; Weyn et al., 2020). Together with an hourly temporal resolution of the video prediction model which aligns with the temporal resolution of operational NWP model output, our application focuses on predicting the diurnal cycle of 2 m temperature. This approach comprises two potential challenges for deep neural networks, that are quick error accumulation in an autoregressive forecast task (see, e.g. Rasp et al., 2020) and the prediction of quasi-periodic processes for which deep neural networks are known to struggle with (Ziyin et al., 2020).

Weather forecasting shares some similarities with video prediction by deep learning. Both explore spatio-temporal patterns from previously observed data to generate a plausible future state of the system. Nevertheless, there are at least two main differences. First, video prediction is mostly used for human pose, physical object and trajectory forecasting, where individual objects are often clearly separable from the background and do not interact with each other on several spatio-temporal scales. Temporal patterns are learned from the movement of objects to then generate a series of frames anticipating how a scene might evolve during the next few seconds. In contrast, weather data does not contain clearly separable objects and the physical laws governing the evolution of weather patterns over time are much more complex due to multi-scale interactions (see, e.g., Orlanski, 1975). For instance, a convective system is driven by large-scale flow patterns (e.g. embedded in a synoptic-scale low) and is subject to turbulence processes in the planetary boundary layer (e.g. convection triggering). Vice versa, the convection itself vents the planetary boundary layer and also modifies the large-scale atmospheric state. Due to the multi-scale interactions, the degree of inherent uncertainty in weather predictions is enormous (e.g. Lorenz, 1969). Second, video predictions mostly aim for perceptually realistic looking scenes. Several evaluation metrics such as the peak signal to noise ratio (Mathieu et al.,

2015) or the structural similarity index (Wang et al., 2004) are applied for this purpose in the computer vision domain. However, the degree of physical realism is barely obtainable from a graphical display of weather, for example in weather charts. Due to this difference, meteorologists have developed a broad range of evaluation metrics with careful consideration of their statistical properties (see, e.g., Wilks, 2011) and deep learning must eventually show that it can compete with numerical models according to the same evaluation standards (e.g. Rasp and Lerch, 2018; Leinonen et al., 2020).

The application of deep neural networks in weather and climate science is still in its infant stage. While some studies experimented with emulators of physical parametrizations within atmospheric models (Brenowitz et al., 2020; Chantry et al., 2021) or processed direct model output for improved forecast products (e.g. Sha et al., 2020; Grönquist et al., 2021), others directly explored video prediction approaches for weather forecasting. So far, relative simple architectures such as fully convolutional u-shaped encoder-decoder networks (U-Net) or convolutional layers coupled with Long Short-Term Memory (LSTM) cells (so called ConvLSTM) are commonly used in the weather forecast domain (e.g. Kim et al., 2017; Weyn et al., 2019; Wang et al., 2021b). In parallel, the performance of deep learning models for computer vision tasks has continuously improved with increased complexity and more refined concepts of the neural network architectures. Since the breakthrough of AlexNet (Krizhevsky et al., 2012) in the ImageNet challenge (Deng et al., 2009), convolutional neural networks and their variants have seen rapid development (e.g. Shi et al., 2015; Canziani et al., 2016). Recently, Generative Adversarial Networks (GAN; Goodfellow et al., 2014), Variational Auto Encoders (VAE; Kingma and Welling, 2013), and vision transformer networks (Dosovitskiy et al., 2020; Caron et al., 2021) have become increasing popular and are nowadays combined with previous approaches to further improve on machine learning benchmark datasets (see Oprea et al., 2020, for a review).

While data-driven neural networks are continuously improved in computer vision, there is recently growing interest in applying physics-informed neural network (PINNs). PINNs aim to leverage the power of neural networks as universal function approximators by explicitly encoding the underlying physical laws expressed in partial differential equations (Raissi et al., 2019) and therefore constitute a promising framework for atmospheric dynamics described by the Navier-Stokes equations, the continuity equation for moist air and the first law of thermodynamics. However, PINNs have only been applied to highly simplified versions of the Navier-Stokes equation yet (e.g. Rao et al., 2020; Jin et al., 2021) and furthermore may suffer from severe convergence and accuracy problems for processes on multiple spatio-temporal scales (Fuks and Tchelepi, 2020; Raissi et al., 2020; Wang et al., 2021a) such as the real atmosphere.

Due to the existing fundamental challenges in applying PINNs to real-world meteorological problems, we focus on data-driven neural networks in this paper. Particulary, we explore to what extent such more advanced deep learning models with the capability of capturing non-linear relations in the data provide opportunities to enhance the predictive skills of machine learning in Earth Science applications. Accordingly, we have applied a state-of-the-art DL architecture, namely the stochastic adversarial video prediction (SAVP) model which combines ConvLSTM, GAN and VAE architecture components (Lee et al., 2018), to a simplified meteorological forecast problem and compare its results with those from a ConvLSTM model. For convenience, we make use of data from the ERA5 reanalysis system (Hersbach et al., 2020) provided by the European Center for Medium-Range Weather Forecasts (ECMWF). This data has the big advantage of providing a comprehensive estimate on the atmospheric state without suffering from sparse observational data with varying biases due to different measurement

techniques (e.g. station sites, radiosondes and satellite observations). Besides, the gridded dataset allows for straightforward applications of convolutional operators. Ultimately, a DL forecast system should work directly on the observational data.

Within the scope of this study, we seek to answer the following research questions: 1) How well do video prediction models perform in predicting the diurnal cycle of 2 m temperature? 2) Is there a clear advantage of using more sophisticated DL architectures? 3) How do different components in composite model architectures such as SAVP affect the forecast quality? 4) How sensitive is the model performance with respect to external parameters (spatial domain, additional predictors and training dataset size)?

The manuscript is organized as follows: Section 2 will give a thorough review of the state-of-the-art deep learning models for video prediction and also presents some related work on weather forecasting. Section 3 introduces the meteorological dataset and describes the video prediction models that are deployed in this study. In Section 4, a detailed analysis of the model results is presented based on standard evaluation metrics from the domain of computer vision and from the meteorological community. The effect of the different components in SAVP models are analyzed through the sensitivity analysis for the scaling factors on L1-loss. We also present the results of sensitivity analysis to evaluate the impacts of input variable selection, the size of the spatial domain, and the length of the training dataset. Finally, Section 6 summarizes the findings and provides an outlook on the future avenue of weather forecasting with video prediction methods.

2   State-of-the-art video prediction models

2.1   Deep learning for video prediction

Common machine learning techniques for video prediction can be categorized into Recurrent Neural Networks (RNN) (Oliu et al., 2018; Wang et al., 2018), Adversarial Learning (Goodfellow et al., 2014; Mathieu et al., 2015), and VAE (Patraucean et al., 2015). While different recurrent network architectures have been developed over the last years, LSTM cells combined with convolutional layers as proposed by Shi et al. (2015) have been widely applied for video prediction as baseline model to compare with other state-of-the-art methods (Villegas et al., 2017; Guen and Thome, 2020). The combination of convolution with LSTM enables the DL model to capture spatio-temporal dependencies and thus make predictions about the temporal evolution of spatial patterns, which is the core task of video prediction.

Despite the early success of ConvLSTM models, they are prone to generate blurry images, which don't look very realistic (Denton and Fergus, 2018; Ebert et al., 2017). The reason for this can be attributed to the loss function used in ConvLSTM models where the L1- and L2-loss constitute common choices. These losses measure the point- or pixel-wise distance between prediction and ground truth and rely on the assumption that the data follows a Gaussian distribution. However, L1- and L2-losses perform poorly when the data is drawn from multi-modal distributions or from non-Gaussian distributions. The problem gets worse with growing uncertainty of the future state (Mathieu et al., 2015). This is because the model tends to converge towards the average of all the possible future states on a point-wise level even if the average values themselves have low probability. This failure in capturing and reflecting the statistical nature of the underlying data leads to a rather quick degradation of forecast accuracy with increasing lead times as noted by Mathieu et al. (2015) and Sun et al. (2019).

As an alternative, GAN-based architectures have been developed, which use adversarial loss to learn the underlying data statistics among multiple equally probable modes and therefore mitigate blurriness. GANs constitute a composite model architecture which consists of a generator and a discriminator model. The discriminator is trained to distinguish between real and artificially generated video sequences. Conversely, the generator gets optimized to fool the discriminator, i.e. it aims to produce video sequences that cannot be differentiated from real ones by the discriminator. By training both models adversarially, the generator must learn the statistical properties of the underlying data and thereby becomes capable to generate *perceptually realistic* images (Oprea et al., 2020).

However, GAN models have also their shortcomings. It is well known that these models may lack diversity in the predicted video sequences which is commonly referred as *mode collapse* in the computer vision community (Isola et al., 2017; Lee et al., 2018). Approaches to overcome model collapse are either to optimize on the Earth Mover distance (Wasserstein GANs) instead of the cross-entropy loss (Gulrajani et al., 2017) or to embed a VAE framework (Kingma and Welling, 2013). VAEs, like other likelihood-based models, can play a complementary role to the GANs and generate more dispersed samples, better learn the data distribution and avoid mode-collapse.

To leverage the advantages of different architectures, Lee et al. (2018) proposed a model architecture that tries to overcome the aforementioned shortcomings by combining three different model architectures. Their SAVP approach incorporates VAE and GAN components together with ConvLSTM cells. Since SAVP leverages the advantages of both, this model demonstrates very good forecasting capability when applied to common ML benchmark datasets such as Moving-MNIST, BAIR Push and KTH (Franceschi et al., 2020; Jin et al., 2020).

2.2 Video prediction in weather forecasts

Precipitation nowcasting with a lead time of up to three hours is one of the most common application of video prediction models. Lagrangian persistence approaches with optical flow (Reyniers, 2008; Ayzel et al., 2019) are well established and already outperform NWP models. The limited performance of NWP models for such short-term forecasts is related to spin-up effects after initialization and to the data assimilation procedure, which is challenged by quickly varying atmospheric processes with non-Gaussian statistical properties such as cloud formation and precipitation. However, optical flow methods fail to capture any developments in the precipitation patterns and thus, advanced deep neural network architectures have been recently applied to attain further improvements. Corner stones in the history of precipitation nowcasting are the study by Shi et al. (2015) and the development of PredRNN (Wang et al., 2017) who both applied ConvLSTM models for this task. Recently, different model architectures with increasing complexity have been tested such as attention models (e.g., Sønderby et al., 2020) and deep U-Nets (e.g Ayzel et al., 2020). Recently, GAN-based models are becoming popular for precipitation nowcasting since they succeed in preserving the underlying statistical distribution and thereby improve in forecasting stronger precipitation events (Liu and Lee, 2020; Ravuri et al., 2021).

For longer lead times, NWP models still constitute the state-of-the-art (Bauer et al., 2015; Schultz et al., 2021), but there have been a few experimental studies which examine the applicability of deep neural networks to generate tailor-made meteorological predictions in the short, medium and seasonal forecast range (more than 6 hours, up to two weeks and beyond).

Weyn et al. (2019) developed weather prediction models using deep convolutional neural networks (CNNs) to predict the 500-hPa geopotential height at a lead time of 14 days. Rasp and Thuerey (2021) proposed a deep residual convolutional neural network (ResNet) to predict global geopotential, temperature and precipitation at 5.625° resolution up to 5 days ahead based on the WeatherBench dataset (Rasp et al., 2020). The study from Weyn et al. (2020) explored a CNN-based model to predict surface temperature patterns. These results and other studies such as Scher (2018); Chattopadhyay et al. (2020) show that basic meteorological features (e.g. evolving Rossby waves) can be predicted from DL models and that a realistic seasonal cycle with prescribed variations in top-of-atmosphere solar forcing can be produced. Even though the DL models still cannot compete with operational NWP models on high spatial resolution, these first results are promising. One aspect which makes DL models particularly attractive is that they are computationally cheap once the neural network has been trained.

However, despite of these initial successes, we observe that DL models for weather forecasting optimized on the L1- and L2-loss also suffer from a similar issue as generating "blurry images" in computer vision tasks. Distinct meteorological features such as precipitation patterns or weather fronts often get smoothed and thus, the predicted meteorological fields exhibit statistical properties that do not match the observed ones.

To improve handling of the inherent uncertainty and to preserve the high spatio-temporal variability in meteorological forecast products, Bihlo (2020) trained conditional GAN (cGAN) models based on the pix2pix architecture (Isola et al., 2017) with a U-net deployed for the generator. With this architecture, he predicted the $500\,\mathrm{hPa}$ geopotential height, the $2\,\mathrm{m}$ temperature and total precipitation for a maximum lead time of 24 hours and attained encouraging results for the two former quantities. Similar to our study, they used ERA5 reanalysis data sliced to a region over Europe and attained promising results on a coarsened $0.5°$ grid.

Our study builds on these recent works by employing the SAVP model architecture to weather forecasts over 12 hours. As described above, SAVP combines the advantages of GANs with those from VAE, and we can thus hope to obtain accurate predictions with sharp features. We compare the SAVP results to a simple ConvLSTM model to probe the sensitivity of the forecast quality on the complexity of the model architecture. Furthermore, we examine the impact of the target domain size, the number of selected predictors and the size of the training dataset.

3 Data and Models

3.1 Dataset

The ERA5 reanalysis dataset provided by the ECMWF is used as the data source in this study (Hersbach et al., 2020). Reanalysis data combines a numerical weather prediction (NWP) model, in this case the Integrated Forecast System (IFS) Cy41r2, with sophisticated data assimilation to retrieve an optimized estimate on the atmospheric state. Global atmospheric reanalysis datasets such as the ERA5 play a substantial role in climate monitoring and are also used over a wide range of other applications in Earth Science, e.g. for hydrological studies (e.g. Tarek et al., 2020) or to track progress in numerical modelling (e.g. Haiden et al., 2021).

The original ERA5 reanalysis data is defined on a reduced N320 Gaussian grid with an approximate horizontal grid spacing of $0.2825°$ ($\Delta x \approx 30\,\text{km}$). Since such highly resolved data fields with 640 grid points in latitude direction and about 1280 grid points in longitude direction near the equator would consume too much memory for the video prediction task, we limit our forecasting task to the region of Central Europe (see Fig. 1) and subset the data accordingly.

The deep learning task of our study is to generate hourly forecasts of the $2\,\text{m}$ temperature over the next 12 hours based on the ERA5 reanalysis fields of the previous 12 hours. The two neural networks used are described in Section 3.3.

In the ERA5 dataset, the $2\,\text{m}$ temperature is not a direct output variable of the underlying IFS model. Instead, it is diagnosed by an empirical interpolation scheme based on the skin temperature and the temperature at the lowest model layer placed $10\,\text{m}$ above the ground (Owens and Hewson, 2018). Both quantities are subject to complex interactions between the surface, the planetary boundary layer and the free troposphere. Over land, the skin temperature is driven by radiation fluxes which undergo a diurnal and seasonal cycle and which are strongly modulated by clouds (Liu et al., 2008). Clouds ~~directly~~ impact the incoming solar (short-wave) radiation and control the long-wave radiation budget which also depends on the temperature of the atmospheric column aloft. Turbulence in the planetary boundary layer further couples the near-surface temperature with the temperature at higher levels in the atmosphere (Garratt, 1994).

While there is a great variety of different variables which drive the underlying processes, we use the $850\,\text{hPa}$ temperature ($T_{850\,\text{hPa}}$) and the total cloud cover (TCC) as additional informative predictors. $T_{850\,\text{hPa}}$ corresponds to the air temperature at a height of approximately $1500\,\text{m}$ above sea level and is commonly used to characterize air masses (Huth, 2002, 2004). This variable has been used in previous $2\,\text{m}$ temperature forecasting studies for statistical post-processing of surface air temperatures by machine learning methods (Casaioli et al., 2003; Eccel et al., 2007). The TCC distills key information on the optical properties of the atmosphere which modulates the incoming solar and the out-going long-wave radiation (Sun et al., 2000; Liu et al., 2008). Thus, both variables are assumed to encode relevant drivers of the $2\,\text{m}$ temperature.

A more systematic variable selection process is not conducted in this study. However, we note that further drivers of the $2\,\text{m}$ temperature can be encoded in a data-driven way from the input data sequence as discussed below.

3.2 Preprocessing

To allow the application of convolutional operations on the data, the data was interpolated onto a regular spherical grid with a spacing of $0.3°$ via the Meteorological Archival and Retrieval System (MARS) in this study. On this grid, the target domain over Europe consists of 92×56 grid points in zonal and meridional direction, respectively. Finally, we restrict the time period to the years 2007–2019 (13 years). In this way, no large climate change signals are involved in the temperature field. The data is originally recorded hourly and further processed as samples to train the deep learning network. Each sample consists of 24 time steps, of which 12 is used as input for the next 12 hours forecasts. While this results in about 8400 samples per year, this choice on the total sequence length allows the models to infer the daytime in a data-driven way. Based on one half of the diurnal cycle of $2\,\text{m}$ temperature as part of the input data, the models have to predict the second half which implies an implicit pattern, and complete encoding of the daytime into the forecasting task. Thus, no explicit information on the daytime is provided to

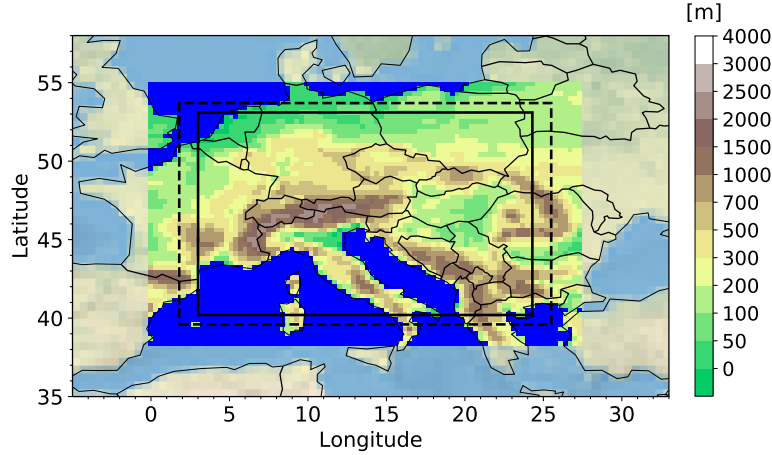



Figure 1. Topographic height of the surface from the ERA5 reanalysis dataset remapped onto a regular, spherical grid with $\Delta x = 0.3°$. The grid boxes of the target domain are highlighted by opaque colours. The dashed and solid lines bound the subdomains in the sensitivity study. The domains comprise 92×56, 80×48 and 72×44 grid points in zonal and meridional direction, respectively.

the models. We furthermore note that no explicit information on the season of the data sequence is provided, but argue that the
230   relevant information can be inferred from the dynamical input data (e.g. temperature of the air mass).

We constructed three cross-validation experiments by selecting different non-overlapping data splits for training, validation and testing (see Fig. 2). Thereby, we make use of 11 years of data for training, while two years are deployed for validation and testing with minimized auto-correlation following the discussion in Schultz et al. (2021). The cross-validation is applied to check the robustness of the trained models over a broad temperature range. By selecting 2010 (CV 3), 2016 (CV 1) and 2019
235   (CV 2) for the testing dataset, we ensure that our trained models are tested on years with relatively cool, on-average and warm temperatures, respectively, within the chosen data period. During training, the validation loss (for tuning the model parameters) operates on data from 2013 (average), 2017 (warm) and 2016 (average).
.

To check how the model performance depends on the spatial domain size and on the provided information of the atmospheric
240   state, we vary the spatial extent of the domain and the number of involved predictors. For the former, smaller target regions are tested, see Fig. 1. The latter is realized by removing TCC from the list of predictors in a first experiment and just inputting T_{2m} in a second experiment. Additionally, we vary the number of training samples by cropping the training dataset and also check the sensitivity on the input sequence length which was discussed to be relevant to encode the daytime in a data-driven way (see above). A comprehensive overview on the sensitivity experiments is provided in Table A1 of the Appendix.
245   Our approach shares some similarities with the study of Bihlo (2020) as we use the same dataset, a similar study region and also focus on short range predictions. However, besides the use of a different neural network architecture (SAVP in our case, cGAN in Bihlo (2020)), there are some distinct differences which make our application potentially more challenging. First, we approximately retain the spatial resolution of the ERA5 reanalysis data on a regular $0.3°$ grid (compared to $0.5°$). Besides,

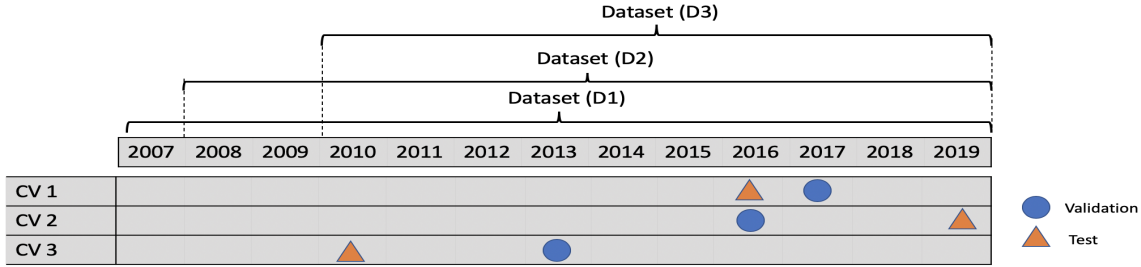



Figure 2. The three cross-validation datasets (CV1, CV2, and CV3) consist of different training, validation and testing samples from the 13 years ERA5 reanalysis data between 2007 and 2019. Each row represents one data partition. The three datasets (D1, D2, and D3) consist of samples from 2007 to 2019, from 2008 to 2019, and from 2010 to 2019, respectively.

the temporal resolution is also considerably higher since we set the time step to one hour compared to three hours. By doing
250 so, our models must learn to represent smaller features of the near surface temperature field and they must better capture the underlying diurnal cycle. By conditioning on the own predictions, errors are expected to accumulate quicker with an hourly time step.

3.3 Model architectures

In the following, we briefly introduce the three video prediction models probed in this study, a simple fully-convolutional
255 neural network (CNN), the convolutional LSTM (ConvLSTM) model and the Stochastic Adversarial Video Prediction (SAVP) model. While a summary on the architectures is provided, more detailed descriptions can be obtained from the original papers, that are the studies by Rasp et al. (2020) on CNN, Shi et al. (2015) on ConvLSTM networks and the work by Lee et al. (2018) for SAVP. The current version of models and code are available and can be accessed from the project website https://gitlab.jsc.fz-juelich.de/esde/machine-learning/ambs/-/tree/GMD1

260 #### 3.3.1 A simple convolutional neural network (CNN)

Following up the study by Rasp et al. (2020), we deploy a simple CNN as one of the baseline models. The CNN consists of 5 layers, and each layer has 64 channels with a kernel size of 5. In contrast to the global forecasts with a spatial resolution of $5.625°$ provided in Rasp et al. (2020), the target domain in our task is restricted to Central Europe and thus, we did not apply periodic convolutions. The mean square error is optimized on the variables of interest ($2\,\mathrm{m}$ temperature, temperature at
265 850hPa, and total cloud cover) for one preceding hour. The model was trained for 20 epochs using the Adam optimizer with a learning rate of 0.0001 and batch size of 4. During forecasting, we use the previous model output as input for the next step, which allows us to obtain forecasts up to 12 hours ahead (iterative forecasting).

3.3.2 The convolutional LSTM (ConvLSTM) model

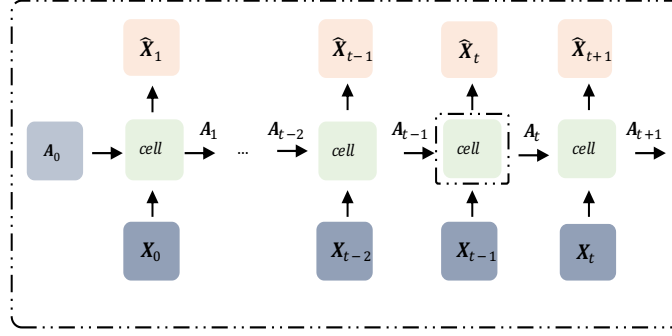
The ConvLSTM model employs convolutional operations which encode the spatial properties of the input data into a hidden state. Temporal coherence is preserved with the help of a gated LSTM, so that an encoded state from all input data is achieved at the end of the input sequence (the encoded atmospheric state over the previous 12 hours in our case). The forecasting network then unfolds this state by conveying both, the LSTM cells' states (cell and hidden state) and the predictions fed in sequentially to come up with a forecast over the next 12 hours. Here, we employ a one-layer ConvLSTM network, i.e. one ConvLSTM layer followed by a 1×1 convolutional layer. A sketch of the model architecture is provided in Fig. 3. We used a batch size of 4 and epochs of 10. The model was trained using mean square error and Adam optimizer (Kingma and Ba, 2014) with a learning rate of $l_r = 0.001$.

3.3.3 The Stochastic Adversarial Video Prediction (SAVP) model

SAVP constitutes a combination of the GAN and the VAE architectures. The model is therefore best described by summarizing both components separately in the following subsections before we explain how these components are coupled together. The explanations are furthermore complemented by a sketch of the SAVP architecture provided in Fig. 4.

Variational autoencoder: The VAE part of SAVP consists of an encoder and a generator. The deep encoder E compresses the information from adjacent time steps into a low-dimensional latent vector $\mathbf{Z}_{t-1} = E(\tilde{\mathbf{X}}_t, \tilde{\mathbf{X}}_{t-1})$ where $\tilde{\mathbf{X}}$ represents either ground data from the input sequence or previously predicted data (i.e. $\tilde{\mathbf{X}}_t \in \{\mathbf{X}_t, \hat{\mathbf{X}}_t\}$). Thus, E encodes the transition between the states at time steps $t-1$ and t into the latent representation $\mathbf{Z}_{t-1}$ which is then fed to the deep generator G for reconstructing the state $\hat{\mathbf{X}}_t$. To control the latent space and to allow random sampling, the learned posterior distribution $q(\mathbf{Z}_{t-1} \mid \tilde{\mathbf{X}}_{t-1:t})$ is kept close to a prior distribution $p(\mathbf{Z}_{t-1})$. Thus, the loss component of the VAE part consists of the L_1-error constituting the reconstruction loss (first term in Eq. 1) and the Kullback-Leibler divergence $D_{KL}\left(q(\mathbf{Z}_{t-1} \mid \tilde{\mathbf{X}}_{t-1:t}) \| p(\mathbf{Z}_{t-1})\right)$ which acts to regularize the posterior distribution of the latent space onto the prior distribution (second term in Eq. 1). The latter constitutes a standard Gaussian distribution whose parameters are optimized with the help of the re-parametrization trick (Kingma and Welling, 2013).

Generative adversarial network: The generator G in SAVP inputs the data from the previous time step $\tilde{\mathbf{X}}_{t-1}$, i.e. either ground truth or previously generated data, to reconstruct the data at time step t. Additionally, the generator is also conditioned on the latent space $\mathbf{Z}_{t-1}$ via sampling. In GAN-based networks the generator G is thereby encouraged to learn the statistical properties of the underlying data. This is achieved by optimizing G to fool a deep discriminator D which is itself optimized to distinguish between *real* data (i.e. the ground truth) and generated data. The loss function of the GAN $\mathcal{L}_{GAN}$ in SAVP applies the binary cross-entropy loss for an adversarial *minimax*-optimization. While the generator tries to minimize $\mathcal{L}_{GAN}$, the discriminator aims to maximize $\mathcal{L}_{GAN}$.

Combining VAE and GAN: For the SAVP architecture, one generator G is set-up which is shared between the VAE and GAN parts. However, two separate discriminators are used which are equivalent in terms of the architecture, but differ in their trainable model parameters. The latter difference arises from the latent embedding that is fed to the shared generator G. For

(a) LSTM Future Predictor Model

(b) ConvLSTM cell

Input   Deep network   Network output

⊗ Tensor product   ⊙ Element-wise multiply

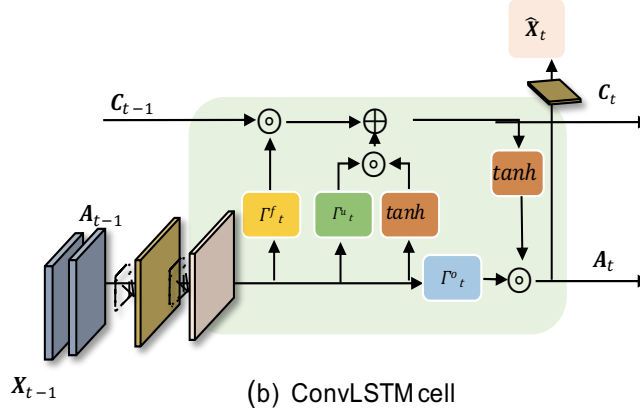
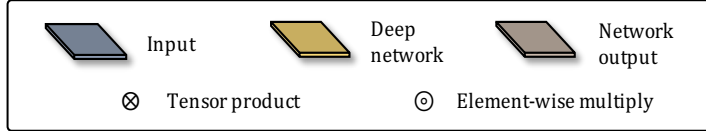



Figure 3. Sketch of ConvLSTM architecture (Shi et al., 2015).

the discriminator related to the VAE part D^{VAE}, $\mathbf{Z}_{t-1}$ is sampled from the posterior distribution $q(\mathbf{Z}_{t-1} \mid \tilde{\mathbf{X}}_{t-1:t})$, whereas sampling from the prior distribution $p(\mathbf{Z}_{t-1})$ is performed for the discriminator of the GAN part. Consequently, two GAN-loss terms become part of SAVP's objective function (see Eq. 1) whose third and fourth term are related to the GAN and VAE component, respectively. The total SAVP loss is calculated as:

305
$$G^*, E^* = arg\min_{G,E} \max_{D,D^{VAE}} \lambda_1 \mathcal{L}_1(G,E) + \lambda_{KL}\mathcal{L}_{KL}(E) + \mathcal{L}_{GAN}(G,D) + \mathcal{L}_{GAN}^{VAE}(G,E,D^{VAE}). \tag{1}$$

The discriminator architecture used by SAVP involves several convolutional layers that operate on the complete data sequence, followed by a fully connected layer. 3D convolutions are applied to handle the spatial and the temporal dimensions simultaneously. The shared generator G involves several ConvLSTM-layers with internal skip connections. These are followed by

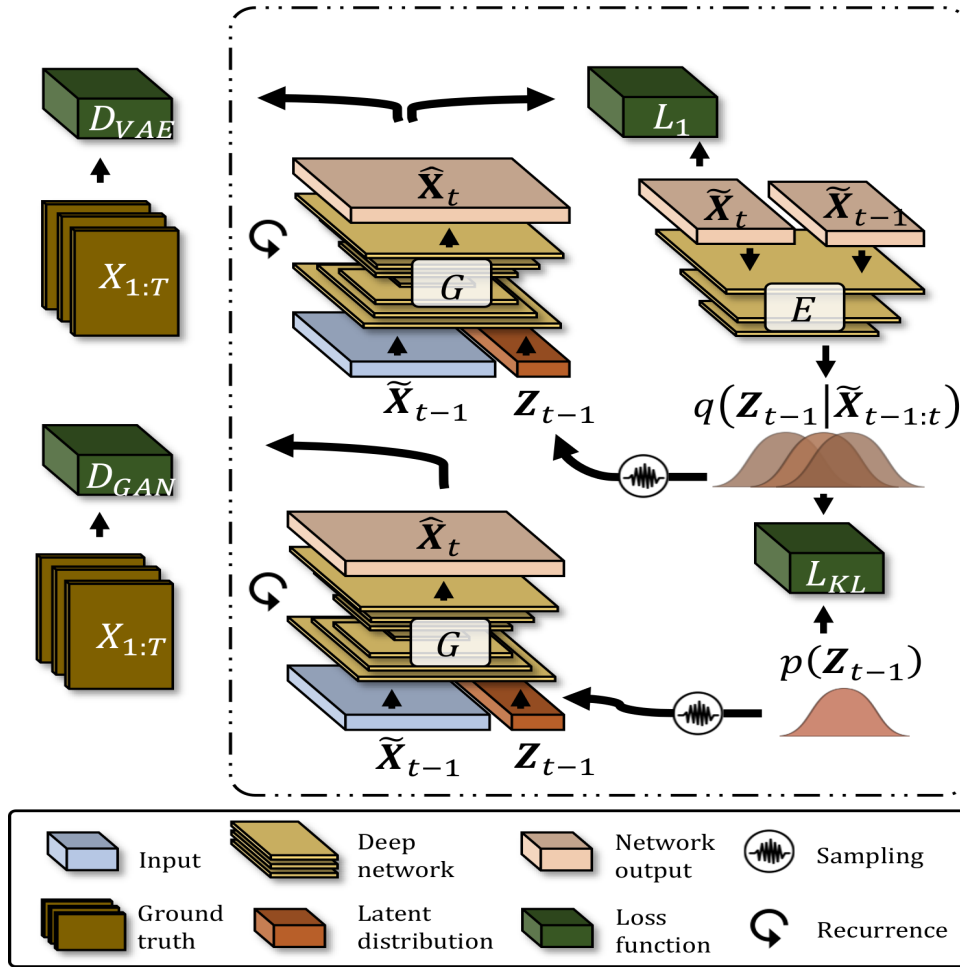



Figure 4. Illustration of the SAVP architecture (from Lee et al., 2018).

two convolutional layers to predict the state at the next time step. A separate composite mask is used to identify features from the data that are displaced between the time steps (e.g. a cold-front in our meteorological application). In total, the number of trainable model parameters sums up to about $14\,\mathrm{M}$. More details on the architecture are provided in Lee et al. (2018).

The default hyper parameters and training schedule in this study have been modified from the original description in Lee et al. (2018). During training, the initial learning rate used with the Adam optimizer is set to $l_r = 2 \times 10^{-4}$. The training applies linear learning rate decay to $l_r = 2 \times 10^{-8}$ between the iteration steps 3000 to 9000. With a batch size of 32, this roughly corresponds to the start of the second epoch and the end of the third epoch, respectively. Overall, the model is optimized on four epochs. For the scaling factors of the different loss components, we choose $\lambda_1 = 10000$ and $\lambda_{KL} = 0.01$. Note that variations in λ_1 attain special attention in our sensitivity analysis below. The scaling factor for the distance in features space is setup to 0.001.

Furthermore, the reconstruction loss only accounts for $T_{2\mathrm{m}}$ which differs from common computer vision applications where all channels enter this loss term.

3.4 Evaluation metrics

For evaluating the video frame prediction models introduced above, we make use of metrics that are commonly applied in the meteorological domain as suggested by Rasp and Thuerey (2021). In particular, we calculate the mean square error (MSE) and the anomaly correlation coefficient (ACC) for the predicted $T_{2\mathrm{m}}$-fields (cf. Appendix Eq. B1 and Eq. B2). While the MSE measures the mean squared distance between the predicted and the analyzed (ground truth) temperature field, the ACC quantifies the agreement in the spatial patterns of departures from the climatological mean. Thus, the ACC is a positively oriented score with a perfect value of 1, whereas the MSE is negatively oriented with a perfect value of 0. In this study, we make use of the uncentered ACC (see Eq. B2) and calculate the climatological mean based on 30 years of data (1991-2020) provided by the ERA5 dataset. The climatological mean is computed for each month of the year and each hour of the day separately. This ensures that the seasonal and diurnal cycle of the near-surface temperature are incorporated.

In addition to the meteorological evaluation metrics, we also choose the Structural Similarity Index (SSIM) which is commonly applied in video prediction to access the perceptual similarity between images (Wang et al., 2004). Transferring to this application, the SSIM quantifies and compares the mean as well as the spatial variability in the predicted $2\,\mathrm{m}$ temperature field against the ground truth and also accounts for covariances (cf. Eq. B7). Although being an evaluation metric from computer vision, it is considered to provide useful information on the forecast quality. Like the ACC, the SSIM is a positively oriented score with a perfect value of 1.

To evaluate the truthfulness of the predicted spatial variability, we also compute the domain-averaged amplitude of the horizontal $T_{2\mathrm{m}}$-gradient. Similar to Sha et al. (2020), we then calculate the ratio r_G of the gradient amplitude from the predictions and the respective ground truth (see Eq. B4). For $r_G(\mathbf{X}, \hat{\mathbf{X}}) = 1$, the predictions share the same local spatial variability as the ERA5 reanalysis data, while $r_G(\mathbf{X}, \hat{\mathbf{X}}) < 1$ ($r_G(\mathbf{X}, \hat{\mathbf{X}}) > 1$) indicates that the local spatial variability is underestimated (overestimated) in the predictions. This metric is similar to the sharpness measure introduced in Mathieu et al. (2015), but takes the Earth's curvature into account and does not scale to the maximal gradient amplitude.

While further details on all evaluation metrics are provided in Appendix C, we verify our models against the persistence forecast in terms of skill scores for convenience. This allows a direct comparison with a reference model which attains a value of S_{ref} for the respective score S. As reference model, we use a simple persistence model which is based on the assumption that today's weather is the same as yesterday's, i.e. the temperature field from the last day is simply copied. Together with the perfect score value S_{per}, the skill score SS_S reads

$$SS_S = \frac{S_m - S_{ref}}{S_{per} - S_{ref}}, \tag{2}$$

where S_m denotes the score value of the considered model forecast.

4 Results

In the following, we evaluate the predictive skill of the SAVP and ConvLSTM models for 2 m temperature predictions up to a lead time of 12 hours. For the presented model results, our default hyper parameters of both models have been tuned to yield best results in terms of the MSE. However, we also provide an ablation study on the L1-loss component in the SAVP model and also probe the sensitivity with respect to input variables, regions and size of training dataset.

4.1 Model performance analysis

The skill scores in terms of the MSE (a), the ACC (b) and the SSIM (c) are displayed in Fig. 5. The uncertainty estimates depicted by the boxes and whiskers are derived through block bootstrapping with a block length of seven days (Efron and Tibshirani, 1994). It is seen that all video prediction models outperform the persistence forecast significantly over the the complete prediction period. The constant MSE of the persistence forecast ($MSE(Persistence) \simeq 7 \, \mathrm{K}^2$) is reduced by about 50% for ConvLSTM and about 70% for SAVP model over the prediction period. Likewise, both models also clearly provide better forecasts of local temperature anomalies from the climatological mean ($ACC(Persistence) \simeq 0.67$) as seen from Fig. 5. Only with respect to the SSIM ($SSIM(Persistence) \simeq 0.66$), the ConvLSTM model loses its forecast skill after a lead time of seven hours (Fig. 5c). In many aspects, the SAVP models also performs clearly better than the ConvLSTM model: The MSE skill score degrades linearly and at a much smaller rate, especially over the first six hours. After 12 hours, the MSE of SAVP tracks at about $2.3 \, \mathrm{K}^2$, while the ConvLSTM-model shows up with a MSE slightly above $3.6 \, \mathrm{K}^2$ (absolute score values are displayed in Fig. A2 of the Appendix). Similarly, the ACC and the SSIM remain closer to 1 for the SAVP model ($ACC(SAVP) \simeq 0.87 > ACC(ConvLSTM) \simeq 0.80$ and $SSIM(SAVP) \simeq 0.73 > SSIM(ConvLSTM) \simeq 0.67$). Thus, as expected, the more complex SAVP model can learn a better representation of the atmospheric state and therefore produces a better prediction of the diurnal cycle of the 2 m temperature compared to ConvLSTM.

However, even though the *global* variability as expressed by the SSIM is better captured with SAVP, the *local* spatial variability is scarcely better than in the ConvLSTM model. In terms of the horizontal gradient ratio, the forecasts of both model degrade continuously with increasing lead time yielding a noticeably underestimation by the end of the forecast period (see Fig. 5d). Thus, the predicted 2 m temperature fields of both video prediction models become too smooth indicating that small-scale variations due to the underlying topography (mountain ranges) and surface type (costal regions) get blurred.

Furthermore, we compare the performance of the video prediction models against the short-range forecasts provided with the ERA5 dataset. These forecasts are initiated at 06 and 18 UTC with a maximum lead time of 18 hours. Since the changes in the assimilation window at 09 and 21 UTC of introduce a systematic shift with respect to the reference reanalysis data, presented scores are limited to lead times between forecast hour 6 and 12. It's seen that there is still a significant gap between the data-driven approaches and the contemporary NWP models which are driven by the fundamental laws of physics. The skill scores of the ERA5 short-range forecasts in terms of the proposed evaluation metrics are higher than the video prediction model and closer to the best score 1.

In addition, we evaluate the performance of the CNN with iterative forecasting from Rasp et al. (2020). The results reveal that the iterative CNN forecasts can only beat the persistence forecasts up to four hours lead time. After that, the model error accumulates quickly and even becomes highly unstable after forecast hour 10 where the MSE becomes as large as $100\,\mathrm{K}^2$ (not shown). Thus, it is evident that the CNN performs considerably worse than the simple ConvLSTM-model which shows that pure CNN models fail to capture the temporal dependency of the data and to obtain skillful forecasts for longer lead times. Recurrent layers are required to transit temporal information which in turn is highly relevant to stabilize the model's long-term forecasting performance.

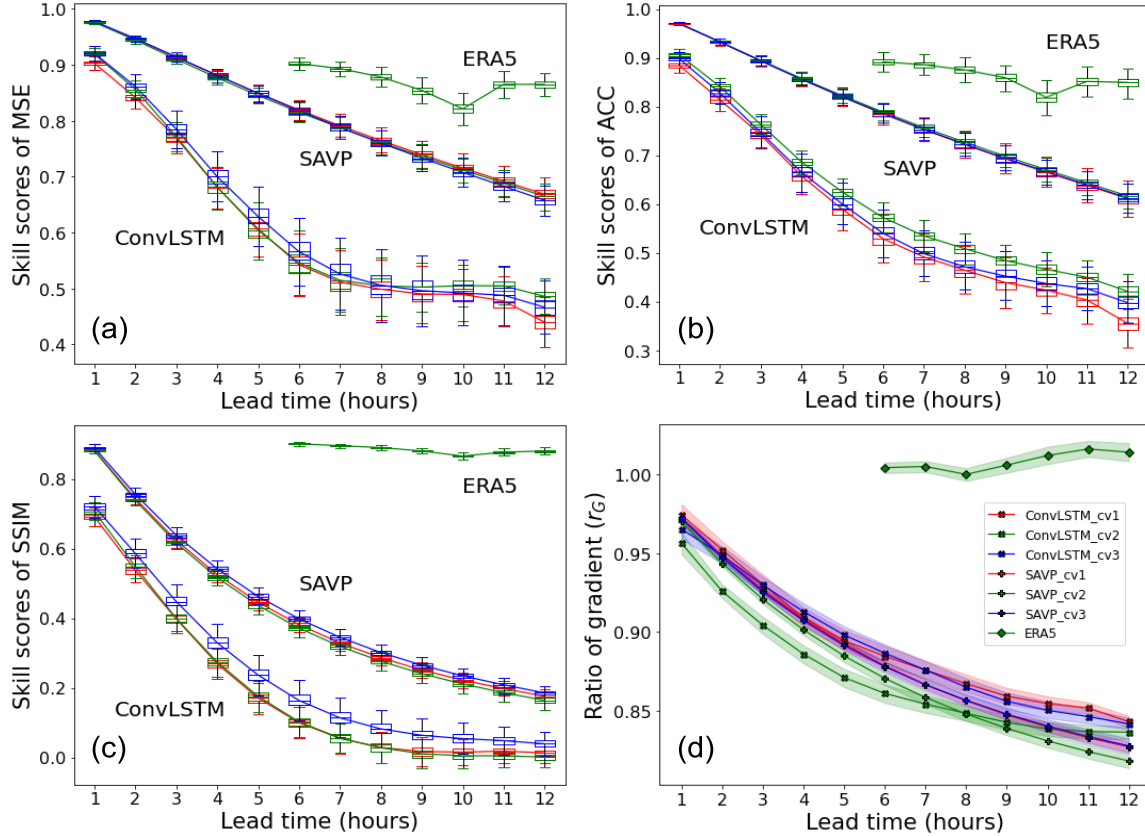



Figure 5. Averaged skill scores of (a) MSE, (b) ACC and (c) SSIM across lead time (x-axis) for the SAVP and the ConvLSTM model, as well as ERA5 short-range forecasts. Both video prediction models are trained with the three cross-validation datasets (with CV 1-3 in D1) displayed in Fig. 2. The testing period of the ERA5 short-range forecasts covers the year 2019 which corresponds to CV2. The persistence serves as the reference forecast. Sampling uncertainty estimated via block bootstrapping is illustrated by box-whiskers showing the inter-quartile and -decile range of the skill score values. (d) shows the ratio of the spatially averaged 2 m temperature gradient for the same video prediction models and ERA5 short-range forecasts. The persistence forecasts (not shown) attain a constant value of $r_G = 1$ since local spatial variability is retained.

To illustrate concretely our statistical findings, we show a representative case study of ConvLSTM and SAVP forecasts starting on 2019-08-02, 03:00 UTC. The first row in Fig. 6 shows the 2 m temperature field from the ERA5 reanalysis (ground truth) for a lead time of six and 12 hours. The differences of the respective ConvLSTM and SAVP model forecasts are presented in the second and third row, respectively, with positive values corresponding to a warm bias.

Apart from growing differences to the ground truth with increasing lead time where SAVP exhibits smaller errors on average, both models forecast strong warming over continental areas. Thus, aspects of the diurnal cycle are captured by the video prediction models. However, it is also noted that the forecast accuracy especially deteriorates around the Alpine region, indicating that both models have problems in predicting the temperature evolution in this area. Besides, differences appear to be bounded by the coastal line with dipole structures visible in the Mediterranean region. This indicates that strong spatial gradients in 2 m temperature tend to be blurred in accordance with the findings in Fig. 5d.

Further insight into the statistical properties of the forecast with respect to the ERA5 reanalysis (the ground truth) can be obtained from conditional quantile plots. These plots visualize important aspects of the joint distribution of forecast and reference data for continuous variables by factorizing it into a conditional and a marginal distribution (Murphy and Winkler, 1987; Wilks, 2011). Figure 7 shows the full joint distribution in terms of the calibration-refinement factorization for lead times of 6 and 12 hours. The median as well as the interdecile range of the ground truth conditioned on the forecasts are displayed by the solid and dashed line, respectively. The histogram at log-scale illustrates the marginal distribution of the respective model forecast.

The central parts of the temperature range are well calibrated in both models, but the ConvLSTM-model shows a broader interdecile range in accordance with the larger MSE. Larger deviations from the 1:1-line are obtained near the tails of the conditional distributions in all four panels. Thus, both models have problems to issue calibrated forecasts when the 2 m temperature is very high (around 310 K) or very low (around 250 K). It is noteworthy that the marginal distribution of the ConvLSTM model results becomes narrower for longer lead times since no temperatures below 252 K (10 out of 8471 samples) are predicted. By contrast, the SAVP-model predicts up to 4 K colder temperatures even for forecast hour 12, although the forecasts are not well calibrated at the lower tail of the conditional distribution. A similar result, but with a smaller amplitude, can also be deduced at the upper tail of the conditional distributions. Thus, the SAVP-forecasts exhibit a slightly higher degree of refinement, also termed sharpness in statistics (Wilks, 2011), compared to ConvLSTM.

4.2 Trade-off between sharpness and accuracy

The term "sharpness" has different meanings in the computer vision and meteorological domains. Sharpness in meteorology characterizes the unconditional distribution of the forecasts. A sharp forecast means that the forecasts are frequently enough distinctly different from the climatological value of the predictand. By contrast, sharpness describes the image contrast at the object edges in the computer vision domain. In the following, we discuss sharpness in this latter sense and analyze the local spatial variability of the 2 m temperature fields in terms of the gradient ratio.

Sensitivity tests were performed by varying the L1-loss scaling factor λ_1 in Eq. (1). While the results are rather insensitive for $\lambda_1 > 100$ (not shown), we notice a stronger dependency of the model performance for smaller values of λ_1. The image

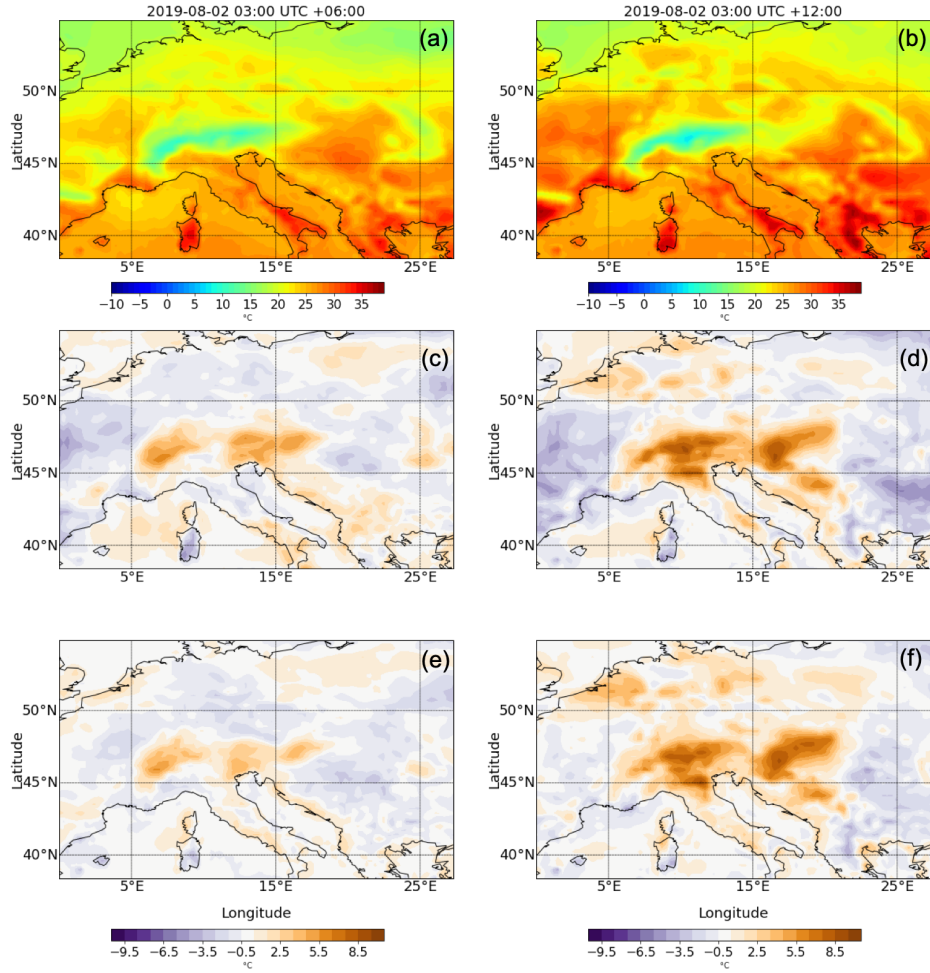



Figure 6. Example forecast for the 2 m temperature with a lead time of six (first column) and 12 hours (second column). (a,b) The ground truth from the ERA5 reanalysis dataset. (c-f) The difference between the forecasts generated by the ConvLSTM (c,d) and SAVP (e,f) where positive (negative) values represent a warm (cold) bias in the forecasts. The initial time for both model forecasts is 2019-08-02, 03:00 UTC. Further time steps of the 12-hour predictions are provided in Appendix A6.

sharpness is improved particularly for longer predictions for smaller scaling factors of the L1-loss, while the MSE is slightly increased (see Fig. 8). This implies that the GAN component in SAVP is largely responsible for maintaining the feature contours. By reducing the strong weight of reconstruction loss, the SAVP model is encouraged to produce temperature fields with a higher local variability, although the errors at grid-point level become larger then. Note that such a trade-off between sharpness in terms of the gradient ratio and accuracy in terms of MSE is observed in weather forecast applications as in other computer vision applications (Lee et al., 2018).

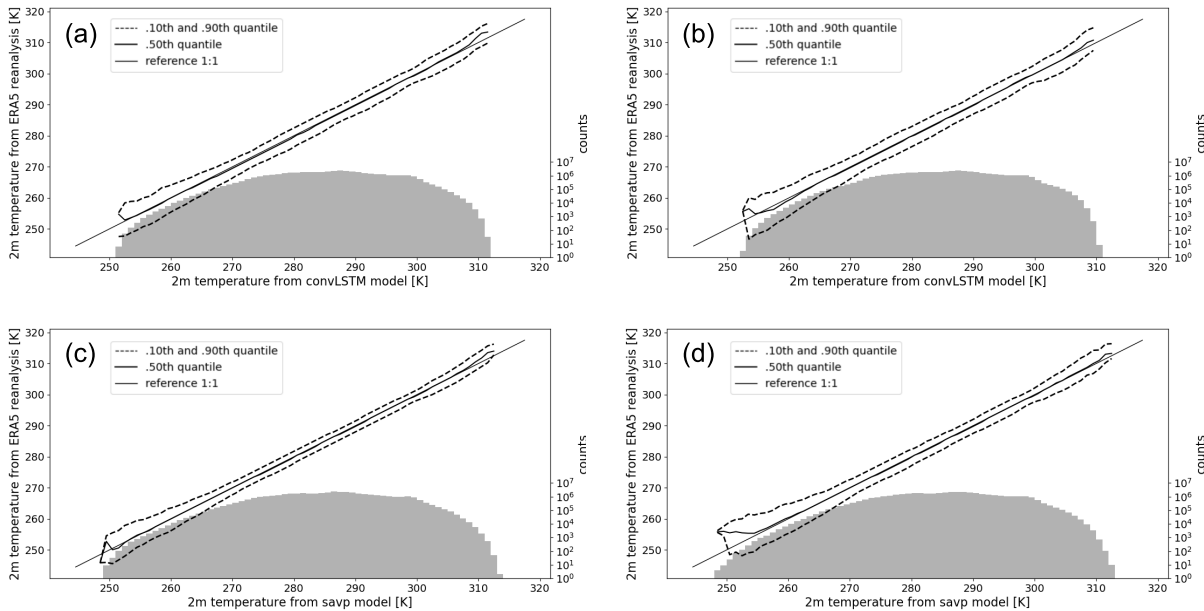



Figure 7. Conditional quantile plots in terms of the calibration-refinement factorization for 2 m temperature forecasts with a lead time of 6 hours (first column) and 12 hours (second column). (a,b) show the factorization of ConvLSTM forecasts, while (c,d) correspond to the SAVP model. The solid straight line denotes the 1:1 reference line of a hypothetical perfect model. The dash lines represent the 10th and 90th quantiles and the bold solid line represent the median of the ground truth data conditioned on the forecasts, respectively. The marginal distribution of the model forecasts is presented as log histogram (right axis, light grey bars).

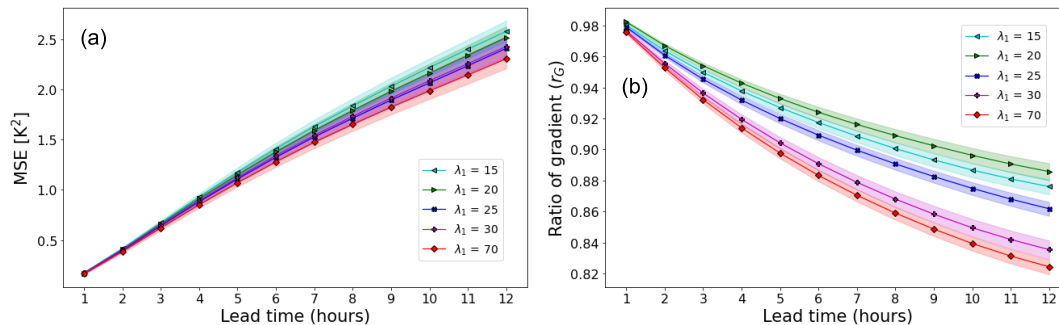



Figure 8. Performance of the SAVP model in terms of (a) MSE and (b) the gradient ratio r_G with variations of the scaling factor for the L1-loss λ_1. Note that the results for $\lambda_1 = 70$ only differ marginally from using $\lambda_1 = 10000$.

This sensitivity study also explains why the SAVP in its original configuration does not outperform the ConvLSTM-model in terms of the gradient ratio (Fig. 5). Due to the very large value of the L1 scaling factor ($\lambda_1 = 10000$), the model is encouraged to optimize for the pixel-wise loss over other losses (e.g. the adversarial loss). Thus, the out-performance in terms of MSE,

430

ACC and SSIM can be attributed to the more sophisticated generator architecture in SAVP alone rather than the adversarial optimization.

In order to gain additional insight into the performance of the GAN component in SAVP on the tails of the $2\,\text{m}$ temperature distribution, two conditional quantile plots generated with $\lambda_1 = 10000$ and $\lambda_1 = 15$, respectively, are provided in Fig. 9. While there are no significant differences for large parts of the conditional distribution, we observe that the median gets closer to the 1:1 reference line at the lower tail of the PDF for $\lambda_1 = 15$. Thus, lowering λ_1 also yields better calibrated model forecasts for very cold temperatures. Note that the lowest temperatures occur in the Alpine region for our target region where grid points are located more than $2000\,\text{m}$ above sea level (see Figure 1). Since the surface elevation varies quite strongly over the mountainous region, the preservation of large local temperature gradients due to the underlying topography is a necessary prerequisite for well-calibrated forecasts in this region.

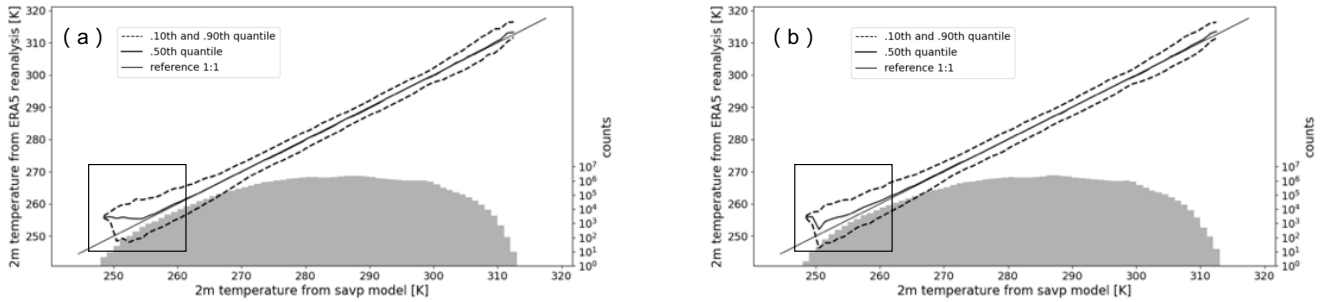



Figure 9. Conditional quantile plots in terms of the calibration-refinement factorization for $2\,\text{m}$ temperature forecasts with a lead time of 12 hours of the SAVP model with (a) $\lambda_1 = 10000$ and (b) $\lambda_1 = 15$, respectively.

4.3 Sensitivity analysis

In the following, we describe further sensitivity tests on the domain size, the selected predictor variables, and the amount of training data of the SAVP model (Experiments 2, 3, and 4, c.f. Tab. A2 in the Appendix). The analysis thereby focuses on evaluating the model performance in terms of the MSE. However, the results of the other evaluation metrics are also briefly presented and the corresponding plots are attached to the Appendix C.

The models in Experiment 2 are trained on data from all 11 years, but the domain size is reduced from 92×56 to 80×48 and 72×44. Note that the evaluation for this experiment is conducted on the inner 72×44 domain to allow for direct comparison. It is seen that the performance in terms of the MSE slightly deteriorates when the target domain becomes smaller (Fig. 10a). Interestingly, the reduction in MSE is most pronounced when enlarging the target domain from 80×48 to 92×56 (reduction in MSE by about $0.2\,\text{K}^2$). Similar results are obtained in terms of the ACC. However, in terms of the SSIM and the gradient ratio r_G only minor changes due to variations in the domain size are observed.

In Experiment 3, the set of predictor variables is reduced gradually from three ($T_{2\text{m}}$, $T_{850\,\text{hPa}}$, TCC) to two ($T_{2\text{m}}$, $T_{850\,\text{hPa}}$) and further to one ($T_{2\text{m}}$). Note that this experiment is conducted on the smaller target domain with 80×48 grid points on which

the average MSE tracks about $0.2\,\mathrm{K}^2$ higher than on the largest domain with 92×56 grid points. While the MSE is fairly insensitive with respect to the inclusion of the TCC, a significant increase is observed when $T_{850\,\mathrm{hPa}}$ is dropped from the list of predictors. For the former case, the MSE is just increased by about $0.1\,\mathrm{K}^2$, whereas the latter results in a MSE of $3.7\,\mathrm{K}^2$ after a lead time of 12 hours. Thus, the temperature of the airmass is more relevant than the total cloud cover for predicting the diurnal cycle of $2\,\mathrm{m}$ temperature in our study. In terms of the other evaluation metrics (ACC, SSIM and r_G), similar results are obtained.

To evaluate the impact of the amount of training samples on the forecasting performance, the size of the training dataset is reduced from 11 years in D1 to 10 years in D2 and eight years in D3, respectively (see Fig. 2). For the sake of a fair comparison, we fixed the validation dataset to 2016 and the testing dataset to 2019, respectively. Removing a single year from the training dataset does not affect the model performance in terms of the MSE (Fig. 10c). When three years of data are dropped from the training dataset, we notice a slight deterioration (MSE increases to about $2.45\,\mathrm{K}^2$ from $2.3\,\mathrm{K}^2$). In terms of the ACC, the SSIM and r_G, the model performance is also rather insensitive against the variations in the size of the dataset probed in this study.

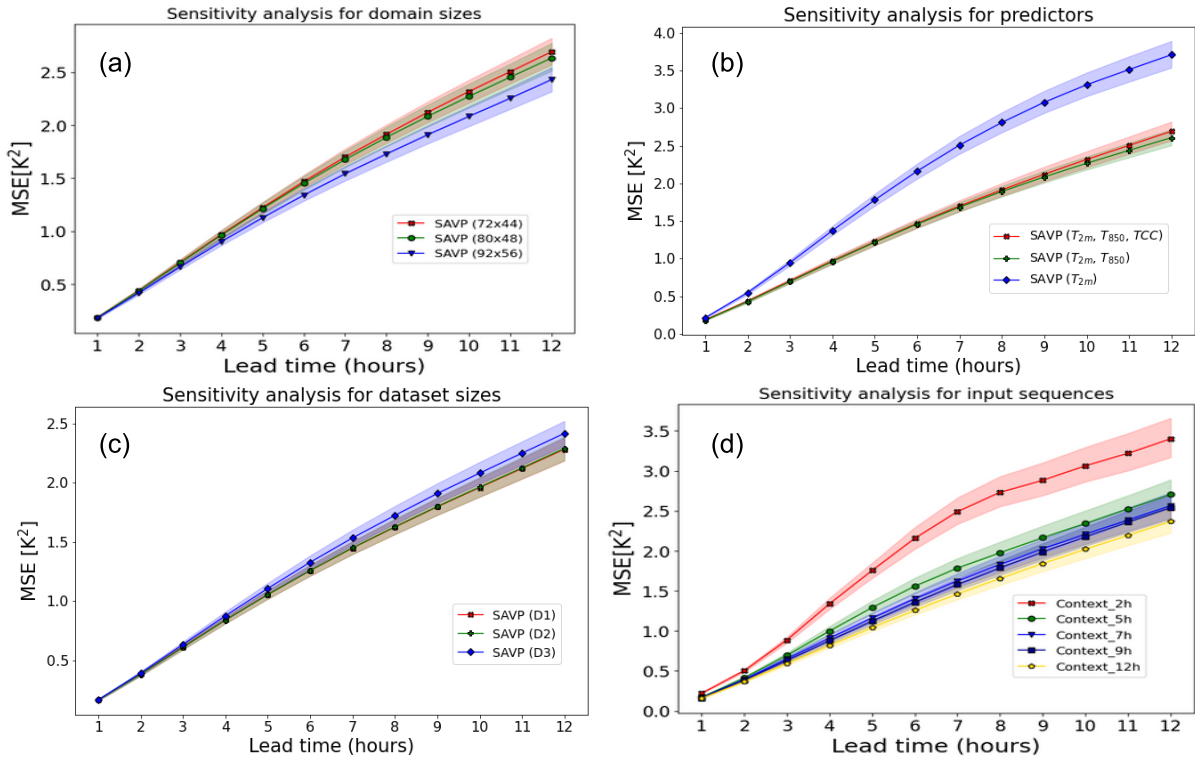



Figure 10. Averaged MSE of $2\,\mathrm{m}$ temperature forecasts with the SAVP model for different sensitive experiments using (a) different domain sizes (red – 92×56, green – 80×48, blue – 72×44), (b) different predictors (red – $T_{2\mathrm{m}}$, $T_{850\,\mathrm{hPa}}$ and TCC, green – $T_{2\mathrm{m}}$, $T_{850\,\mathrm{hPa}}$, blue – $T_{2\mathrm{m}}$) and (c) different sizes of the training dataset (red – 11 years, green – 10 years, blue -8 years), (d) different input sequence lengths. The sampling uncertainty estimated via block bootstrapping is denoted by colour shading.

To identify whether the models could infer the daytime without providing explicit daytime information, we design Experiment 5 where the length of the input sequence is varied. We notice that providing on-half of the diurnal cycle (12 preceding hours) as input yields the best model performance in terms of the MSE (see Fig. 10d). Reducing the length of the input sequence from 12 hours to five hours results into a successive degradation of model performance. Although the increase in MSE is only about $0.3 - -0.4 \, \mathrm{K}^2$ for a lead time of 12 hours, its impact is stronger than removing three years of training data or than removing TCC from the list of predictors.

A further significant degradation takes place when the input sequence is restricted to the two preceding hours. The MSE then becomes similar to the experiment where the 2 m temperature itself was inputted as the predictor. In terms of the ACC and the SSIM, the sensitivity of the results is similar, however for the gradient raio, we don't observe a significant impact of input sequence length.

5 Discussion

The results presented in the previous section demonstrate that video frame prediction models from computer vision attain some predictive skill in forecasting the diurnal cycle of 2 m temperature. We showed that the SAVP model performs significantly better than a simple ConvLSTM model in terms of several evaluation metrics (MSE, ACC and SSIM). This confirms our expectation that more advanced DL can better extract spatio-temporal features from the input sequence to predict the future state, which in turn is beneficial for meteorological applications, even though these models are originally developed for applications in computer vision.

However, the local spatial variability as measured in terms of the gradient ratio is not necessarily improved in our experiments with the SAVP model. Experiments with varied scaling factors of the L1-loss component λ_1 reveal that the strong weight on the pixel-wise reconstruction loss in our basic hyperparameter setting (cf. Section 3.3.3) is responsible for this behaviour. With $\lambda_1 = 10000$, the adversarial part of the SAVP architecture is effectively neglected. Thus, the improvement seen in the evaluation metrics can be attributed to a more advanced generator which incorporates ConvLSTM cells, several convolutional layers along with skip connections and conditioning information on latent code. Reducing λ_1, the accuracy of the model in terms of the MSE slightly decreases, but the local spatial variability becomes much more similar to the ground truth data. In other words, the adversarial training with the GAN-components encourages the model to preserve local features in the 2 m temperature field which can be attributed to spatial variability due to varying characteristics of the Earth's surface. Note that the latter characteristics (e.g. land-sea contrast and surface elevation) has not been explicitly fed to the models, so that their impact needed to be learnt in a data-driven way. Additionally, the forecasts tend to be better calibrated for very cold conditions. Since very cold conditions constitute the tail of the marginal distribution of 2 m temperature, it is hypothesized that the GAN-component in SAVP may help to forecast extreme temperature events.

Further sensitivity experiments reveal that the prediction of the diurnal cycle of 2 m temperature can significantly benefit from incorporating additional predictors. In particular, the temperature at 850 hPa provides additional information on the air mass characteristics which in turn yield a strong reduction of the MSE by at least 30 %. However, adding the total cloud cover

surprisingly barely contributed although cloud drive the energy fluxes at the surface which in turn drive the diurnal cycle of
500 near surface temperature and the planetary boundary layer in general (e.g. von Engeln and Teixeira, 2013; Chepfer et al.,
2019). One reason for this result might be that the model has problems in predicting the future cloud cover since the underlying
microphysical processes are highly complex (see, e.g., Khain et al., 2015) and that we only optimize the model on the 2 m
temperature. While this turned out to be beneficial in our study (not shown), meaningful feature abstraction from quickly
varying predictors becomes challenging.

505 Furthermore, the model performance is slightly improved upon enlarging the target domain. On the one hand side, this might
be attributed to an improved feature abstraction of the large-scale atmospheric conditions (e.g. the advection of air masses).
On the other hand, the synoptic-scale features have limited relevance at sub-daily time scales since the time and spatial scales
of atmospheric processes are correlated (see, e.g. Orlanski, 1975). Besides, the largest domain 92×56 includes the largest
fraction of marine pixels. Since the 2 m temperature exhibits much smaller diurnal variations over the sea surface (Ginzburg
510 et al., 2007), the prediction becomes simpler for these regions which in turn translate to smaller prediction errors (see also
Fig. 6).

In addition, the MSE of the SAVP model trained with 11 years of data is slightly decreased, but compared to the additional
amount of data included (37.5% compared to the 8 years-dataset), the effect is judged to be minor. We argue that the dataset
should probably not include less data than probed here, but contrarily, including more data from the ERA5 reanalysis database
515 is not expected to provide substantial benefits. It could even be that stronger climate change signals may outweigh the added
value of including more data to the dataset.

It is worth to mention that inputting one half of the diurnal cycle of 2 m temperature is beneficial to model's forecasting
capability. Limiting the input sequence to only two hours yields to a strong increase in the MSE by about $1.5 \, \mathrm{K}^2$ which is
equivalent to removing all informative predictors besides to the 2 m temperature itself. Since the performance already deterio-
520 rates for smaller changes to the input sequence, we conclude that the model can infer the daytime from the input sequence in a
data-driven way provided that it covers at least one half of the day.

Our study shares some similarities with the study of Bihlo (2020) which also presents short-range forecasts of the 2 m
temperature over Europe with a GAN-based model. While his predictions attain a fairly low RMSE of about $0.53 \, \mathrm{K}$ for
12-hour forecasts which is considerably smaller than the model performance with our SAVP model ($RMSE(SAVP) =$
525 $\sqrt{MSE(SAVP)} \simeq 1.5 \, \mathrm{K}$), direct comparison is limited due to relevant changes in the target of the forecast product. First,
the spatial resolution of the target product is higher with $0.3°$ compared to $0.5°$ and thus, local spatial variability must be
captured more precisely in our case study. Second, we choose a hourly forecast product to focus explicitly on the predictability
on the diurnal cycle. Thus, twelve consecutive forecasts are required to generate a 12-hour forecast with our SAVP which is
considerably more than four time steps in Bihlo (2020). Thus, the increased temporal resolution of our forecasts come at the
530 price of a stronger error accumulation, since the forecasts are conditioned on the previous hour.

Furthermore, we notice that the gap between data-driven neural networks for meteorological forecasts and contemporary
NWP models is still considerable. While a RMSE of $1.5 \, \mathrm{K}$ for a 12-hour forecast is attained with our SAVP model, contempo-

rary global NWP models show up with a RMSE of about $0.4\,\mathrm{K}$[1]. Meanwhile, they also provide a higher spatial solution around $0.1° \simeq 10\,\mathrm{km}$. However, in light of the long development history of NWP models for several decades (Bauer et al., 2015), the results with data-driven neural networks is already encouraging. Thus, further research as presented in the following section, may further close the gap between deep neural networks and classical NWP models.

6 Conclusion

In this study, we have explored the application of video prediction models, originally developed for computer vision applications, to forecast the sub-daily temperature evolution over Europe. While the results show that more sophisticated model architectures such as the SAVP model and the inclusion of informative predictors such as the $850\,\mathrm{hPa}$ temperature can improve significantly the model performance, we also shed further light on the orchestration of different loss components in the composite SAVP architecture. Tuning the model on the L1-loss optimizes for the MSE, but also leads to a strongly underestimated local spatial variability. Conversely, choosing a smaller weight on the L1-loss leads to slight increase in MSE, but the spatial variability gets better preserved due to a relatively stronger contribution of the GAN-component in SAVP.

The findings in our study and the persisting large gap to NWP models motivates future work which aiming to improve the performance of the underlying deep neural networks. First, one may consider testing further state-of-the-art video prediction models from computer vision which continue to develop at quick momentum. Further advanced GAN-based models (e.g Brock et al., 2019; Clark et al., 2019; Qi et al., 2020) or the recent success in vision transformers (e.g. Caron et al., 2021; Yan et al., 2021) are appealing candidates which may help to reduce the above mentioned gap. Apart from improving the model architecture, our results also suggest that deep neural network can benefit greatly from adding further predictors beside the target variable to further improve the forecast skill. In our case, the $850\,\mathrm{hPa}$ temperature proved to be beneficial for the model performance and it is likely that other dynamic predictors such as surface fluxes or near-surface wind can contribute to the model performance as well. Also static fields such as surface elevation and the land-sea mask should be considered (Sha et al., 2020; Lezama Valdes et al., 2021). Thus, a more systematic predictor selection is an appealing candidate to further improve the forecast skill.

Another way would be to exploit explicitly physical knowledge. This could be realized during preprocessing via feature engineering or during training by formulating physical constraints (de Bézenac et al., 2019; Karniadakis et al., 2021). In some cases, even simple physical constraints can be beneficial and can furthermore increase the realism when predictions beyond the training data space need to be issued (Karpatne et al., 2017).

Additionally, enlarging the forecast domain is considered to be helpful, especially when the lead time is extended. For medium-range forecasts, even a processing of the global atmospheric state, but it is mentioned that this would result into enormous memory requirements on the operating GPU used for training, at least when a highly resolved forecast product is demanded (c.f. Dueben and Bauer, 2018).

---

[1]This value was extracted on 2021-12-10 from the NCEP/EMC global model verification website.

Due to the multi-scale, non-linear interactions in atmospheric processes, the uncertainty in weather predictions tends to be large (see, e.g., Lorenz, 1969) and quantifying this uncertainty is considered to be crucial in meteorology (Vannitsem et al., 2021). The demand for a probabilistic framework further increases when the forecast product involves higher inherent uncertainty than the 2 m temperature. The SAVP model can also be used for probabilistic forecasting by incorporating the stochasticity hrough input noise and the latent space of VAE component. The VAE component encode the input into the latent representation that returns a distribution instead of a single point. The decoder synthesizes the frames by sampling random latent code from this distribution. This approach proved to be effective in generating more diverse samples on machine learning benchmark datasets in computer vision (Lee et al., 2018), but it has to be checked if this also applies to meteorological forecasting tasks.

Precipitation is a typical example for such a meteorological quantity since it is subject to complex interacting microphysical and dynamical processes on small spatio-temporal scales (Sun et al., 2014). While precipitation nowcasting is already gaining momentum in the meteorological community (Prudden et al., 2020), video prediction models may also be helpful to extent the forecast range beyond a few hours. This is motivated by the fact that even contemporary NWP models still have problems in predicting precipitation events, while first DL-based applications already start to compete successfully with these models (see, e.g., Espeholt et al., 2021; Ravuri et al., 2021).

Code availability.

The exact version of the model used to produce the results used in this paper is archived on zenodo https://doi.org/10.5281/zenodo.6308774) under the MIT licence. The scripts can run the model and make the plots for all the simulations presented in this paper.

Data availability. TEXT

A small samples with 1 year ERA5 dataset is published is under under Creative Commons Attribution (CC-BY) Licence. The data can be accessed and downloaded through B2SHARE (http://doi.org/10.23728/b2share.744bbb4e6ee84a09ad368e8d16713118). This dataset can help users to run tests on their local machine or perform quick tests. By using these data, the users must run the end-to-end workflow as described in the README.md file of the code repository.

The experiments described in the paper rely on the 13 years of ERA5 data. We provide the processed data in TFRecords format, which can be accessed through https://datapub.fz-juelich.de/esde/esde-nfs/online_publication/2mT_by_DL/. These data can be used directly in the training process.

Sample availability. TEXT

Appendix A: Experimental setup

As shown in Table A2, we conducted four experiments with different settings, i.e. varying input variables, regions, and number
595   of training samples, to explore the feasibility and robustness capability by DL for 2 m temperature forecasting. We finally
obtain eight datasets as listed in Table A1, which will be adapted for different experiments in Table A2.

Table A1. Overview on the datasets used in this study. Note: D1, D2, D3 correspond to the three datasets where the number of years in the
training dataset is varied (11, 10, and eight years, respectively) as illustrated in Fig. 2.

Dataset ID	Training Dataset	Domain size	input/output length	Variables
1	D1	92×56	12/12	TCC, $T_{850\,hPa}$, T_{2m}
2	D2	92×56	12/12	TCC, $T_{850\,hPa}$, T_{2m}
3	D3	92×56	12/12	TCC, $T_{850\,hPa}$, T_{2m}
4	D2	80×48	12/12	TCC, $T_{850\,hPa}$, T_{2m}
5	D2	72×44	12/12	TCC, $T_{850\,hPa}$, T_{2m}
6	D2	80×48	12/12	$T_{850\,hPa}$, T_{2m}
7	D2	80×48	12/12	T_{2m}
8	D2	92×56	2/12	TCC, $T_{850\,hPa}$, T_{2m}
9	D2	92×56	5/12	TCC, $T_{850\,hPa}$, T_{2m}
10	D2	92×56	7/12	TCC, $T_{850\,hPa}$, T_{2m}
11	D2	92×56	9/12	TCC, $T_{850\,hPa}$, T_{2m}

Table A2. Overview on the experiments in this study.

Experiments IDs	Dataset IDs	Models	Target (Scientific questions to be answered)
1	1	ConvLSTM, SAVP, Persistence	Test and compare models performances
2	1,4,5	SAVP	Test sensitivity to the domains
3	4,6,7	SAVP	Test sensitivity to the variables
4	1,2,3	SAVP	Test sensitivity to the size of training data
5	8,9,10,11,2	SAVP	Test sensitivity to input sequence length
6	1	SAVP	Test sensitivity to the scaling factor λ_1

Appendix B: Evaluation metrics

In this section, we provide some mathematical details on the evaluation metrics used in this study, that are the mean squared error (MSE), the anamoly correlation coefficient (ACC), the Structural Similarity Index (SSMI) and the gradient ratio r_G.

600 **B1 Mean squared error (MSE)**

The MSE measures the squared difference between the model data and the ground truth data. Let x_{ij} and $\hat{x}_{ij}$ constitute data on discrete grid points of the ground truth and the forecasts, respectively, where the grid consists of I and J cellcenter positions in zonal and meridional direction (or the width and height in pixels for images), respectively. With $N = I \times J$, the MSE reads

$$MSE = \frac{1}{N}\sum_{i=1}^{I}\sum_{j=1}^{J}[\hat{x}_{ij} - x_{ij}]^2. \tag{B1}$$

605 **B2 Anomaly correlation coefficient (ACC)**

The ACC quantifies how well the spatial position of anomalies matches between the predicted and the ground truth data (i.e. the ERA5 reanalysis data in our case). The uncentered ACC is given by:

$$ACC = \frac{\sum_{i=1}^{I}\sum_{j=1}^{J}\left[\hat{x}_{ij} - x_{ij}^c\right]\left[x_{ij} - x_{ij}^c\right]}{\left\{\sum_{i=1}^{I}\sum_{j=1}^{J}\left[\hat{x}_{ij} - x_{ij}^c\right]^2 \sum_{i=1}^{I}\sum_{j=1}^{J}\left[x_{ij} - x_{ij}^c\right]^2\right\}^{1/2}}. \tag{B2}$$

Here, x_{ij}^c represents the climatological mean which is inferred at each grid point from the ERA5 reanalysis data between
610 1990 and 2019 in this study. Since the 2 m temperature involves a seasonal and a diurnal cycle, the climatology is calculated separately for each month of the year and each hour of the day, respectively.

B3 Structural Similarity Index (SSIM)

The SSIM constitutes a score metric typically applied in computer vision to measure the similarity between two images (Wang et al., 2004). It measures and compares the structural information between the ground truth and prediction images. The sim-
615 ilarity is thereby quantified in terms of luminance, contrast variance and structure. In case of images with multiple channels (i.e. RGB-images), the calculations are done separately for each channel and averaged afterwards.

  – Luminance: The luminance is measured by averaging the pixels' brightness of the images. Letting $\mu_{\mathbf{X}}$ and $\mu_{\hat{X}}$ denote the averaged brightness of the pixels from the ground truth and the generated image, the respective component l of $SSIM$ is given by

620
$$l = \frac{2\mu_X\mu_{\hat{X}} + C_1}{\mu_X^2 + \mu_{\hat{X}}^2 + C_1} \tag{B3}$$

where C_1 is a constant to avoid divisions by zero or very small numbers. Specifically, we choose $C_1 = (K_1 L)^2$ where $K_1 = 0.01$ and L is the dynamic range of input values. In case that the average brightness of the two images matches, $l = 1$ is obtained.

– Contrast: The contrast is measured by calculating separately the standard deviation of the pixel brightness of each image. Letting σ_X and $\sigma_{\hat{X}}$ denote the standard deviation of the ground truth and the generated image with

$$\sigma_X = \left[\frac{1}{N-1}\sum_{i=1}^{I}\sum_{j=1}^{J}(x_{ij}-\mu_X)^2\right]^{1/2}. \tag{B4}$$

The contrast score component c then reads:

$$c = \frac{2\sigma_X\sigma_{\hat{X}}+C_2}{\sigma_X^2+\sigma_{\hat{X}}^2+C_2}. \tag{B5}$$

C_2 is added in analogy to C_1, but for the contrast score component. Here, $C_2 = (K_2 L)^2$ and $K_1 = 0.03$.

– Structure: The structure is computed with the help of the covariance in pixel-space between the ground truth and the generated image and therefore measures the coherence between the two images in terms of variations around their average brightness. With $\sigma_{X\hat{X}}$ denoting the covariance and using the standard deviations of the ground truth and generated image, that is σ_X and $\sigma_{\hat{X}}$, the structural component s of SSIM is defined by

$$s = \frac{\sigma_{X\hat{X}}+C_3}{\sigma_X\sigma_{\hat{X}}+C_3}, \tag{B6}$$

where $C_3 = C_2/2$ serves as an additional constant, as proposed in the study Wang et al. (2004).

Finally, the SSIM is obtained by merging the different components together with the help of

$$SSIM = l^\alpha \cdot c^\beta \cdot s^\gamma. \tag{B7}$$

Here, α, β and γ are disposal (positive) parameters which control the importance of each score component. Here, we use $\alpha = \beta = \gamma_c = 1$. In our particular translation, the luminence corresponds to a measure which compares the domain-averaged 2 m temperature between prediction and the respective ERA5 reanalysis data. Contrast is equivalent to an comparison between the global variability of T_{2m} per time step, while the structure simply compares the covariance between the predicted and the ground truth 2 m temperature fields.

B4   Gradient ratio

The SSIM does not account explicitly account for local variability in the data. Indeed, the l and c components only evaluate the overall average and the variability. Only s measures how the data varies on a grid point (pixel-wise) level. To further analyze the local spatial variability, we calculate the amplitude of the horizontal 2 m temperature gradient. Sha et al. (2020) follow a similar approach in their temperature downscaling application, but rather deploy the Laplace operator than the gradient operator. Besides, their operator does not account for the curvature on the sphere.

In the geographical coordinate system, the horizontal gradient of the arbitrary quantity ψ reads:

$$\nabla_h\psi = \frac{1}{r_E\cos(\varphi)}\frac{\partial\psi}{\partial\lambda}\mathbf{e}_\lambda + \frac{1}{r_E}\frac{\partial\psi}{\partial\varphi}\mathbf{e}_\varphi. \tag{B8}$$

Here, λ and φ denote the longitude and latitude, while r_E is the (averaged) Earth radius.

On a geographical grid, the amplitude of the continuous gradient operator ∇_h can be discretized with finite differences:

$$G_\psi = |\nabla_h \psi(i,j)| \simeq \frac{1}{r_E} \sqrt{\frac{1}{\cos^2(\varphi(j))}\left(\frac{\psi(i+1,j)-\psi(i-1,j)}{2\Delta\lambda}\right)^2 + \left(\frac{\psi(i,j+1)-\psi(i,j-1)}{2\Delta\varphi}\right)^2}, \tag{B9}$$

where $\Delta\lambda$ and $\Delta\varphi$ denote the grid spacing of the underlying grid. i and j correspond to indices on the grid in zonal and
655   meridional direction, respectively. This yields a second-order accurate discretization on a regular grid (constant spacing in horizontal directions).

Now, let G_X and $G_{\hat{X}}$ denote the average of the absolute horizontal gradient of our target quantity over all grid points on the domain (apart from the lateral boundaries) for the ground truth and the predicted data, respectively. The ratio

$$r_G = \frac{G_{\hat{X}}}{G_X} \tag{B10}$$

660   then measures the averaged local spatial variability amplitude in the predicted field compared to the reference data. For $r_G = 1$, the amplitude of the horizontal gradient is on average the same in the prediction and in the ground truth. For $r_G < 1$ ($r_G > 1$), the horizontal gradient is on average underestimated (overestimated) by the model, indicating that the field is visually too blurried (too sharp) following the discussion in Section 3.

Appendix C: Supplementary figures

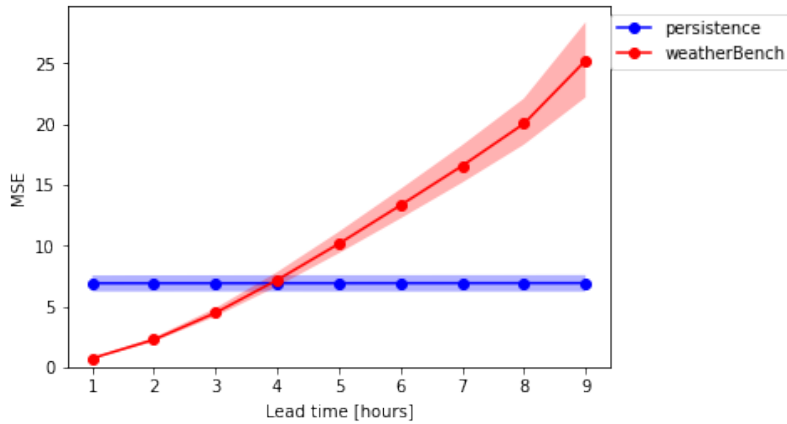



Figure A1. MSE over 9 hours forecast for for 2 m temperature with interdecile bootstrap confidence intervals (shading area) using Weather-Bench convolutional neural network and persistence forecasting.

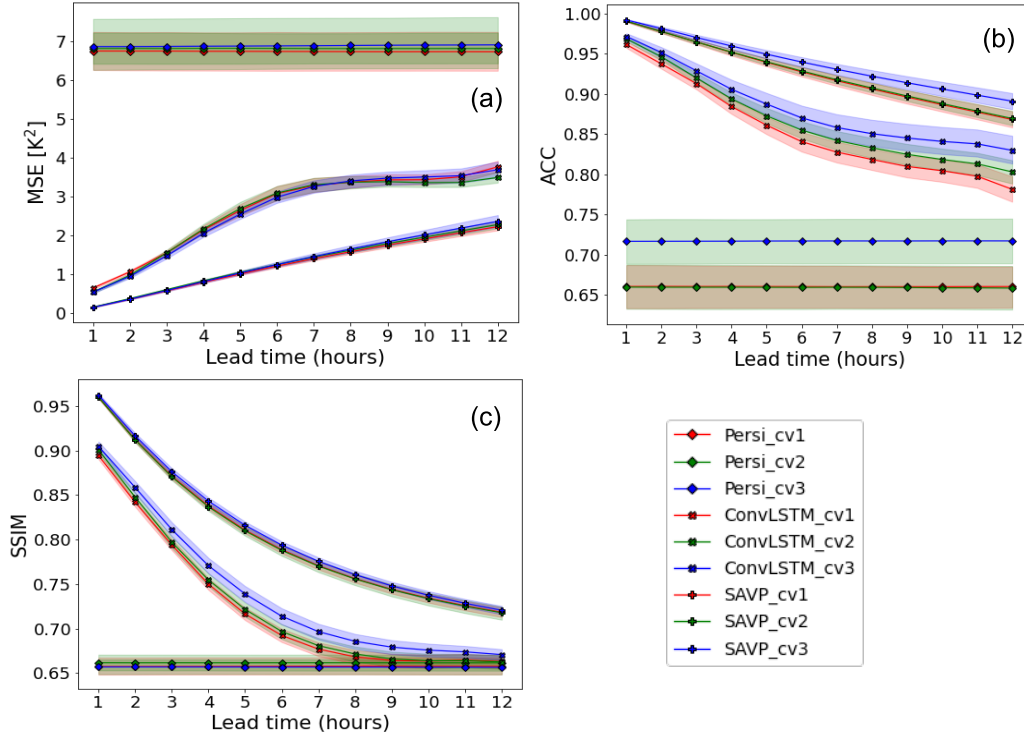



Figure A2. Averaged (a) MSE, (b) ACC and (c) SSIM across lead time (x-axis) for the SAVP, the ConvLSTM and the persistence forecasts from Experiment 1 in Table A2. Like in Fig. 2, sampling uncertainty is estimated via block bootstrapping. However, the interdecile confidence range is now colour shaded for the three cross validation datasets using different models.

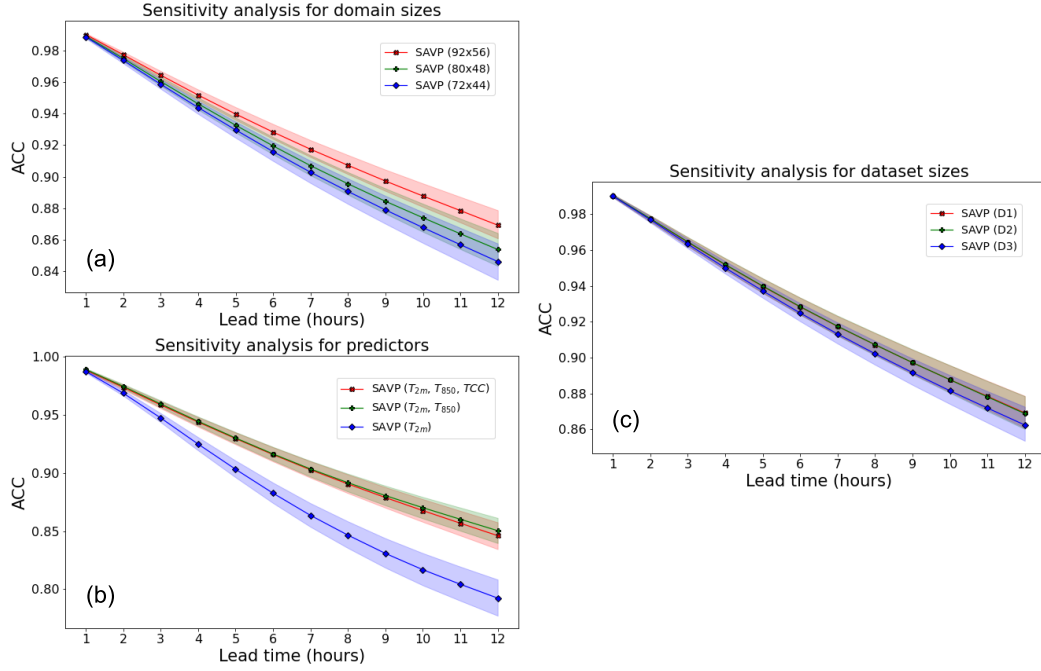



Figure A3. ACC over the forecast period for $2\,\mathrm{m}$ temperature with interdecile bootstrap confidence intervals (shading area) using the SAVP model for different sensitive experiments using (a) different domain sizes (red – 92×56, green – 80×48, blue – 72×44), (b) different predictors (red – $T_{2\mathrm{m}}$, $T_{850\,\mathrm{hPa}}$ and TCC, green – $T_{2\mathrm{m}}$, $T_{850\,\mathrm{hPa}}$, blue – $T_{2\mathrm{m}}$) and (c) different sizes of the training dataset (red – 11 years, green – 10 years, blue -8 years).

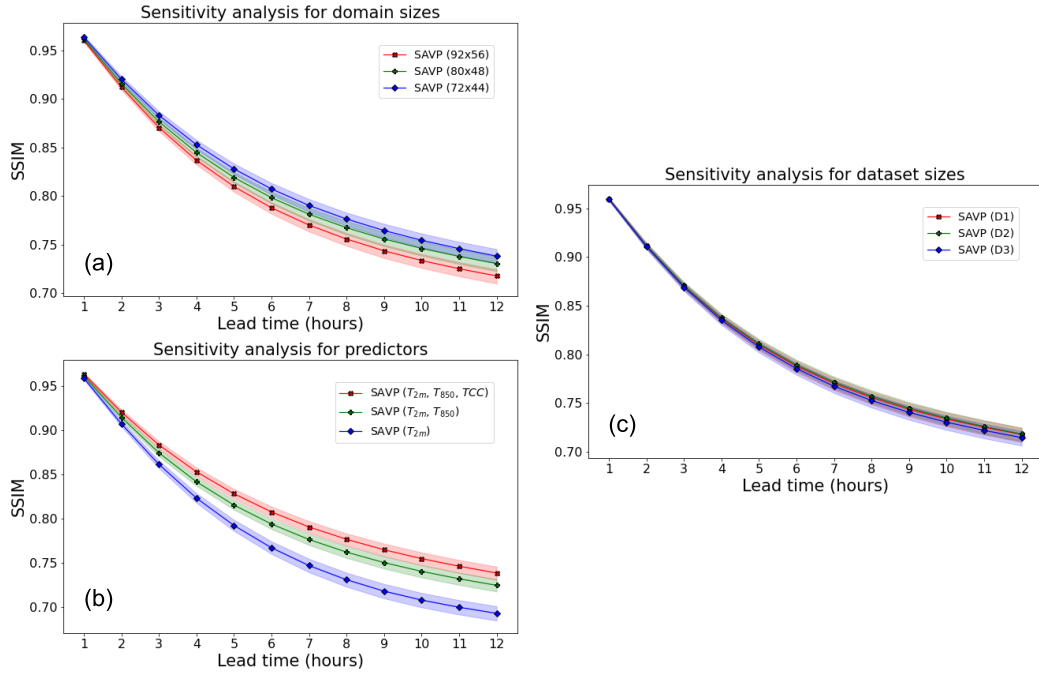



Figure A4. Like Fig. A3, but for SSIM.

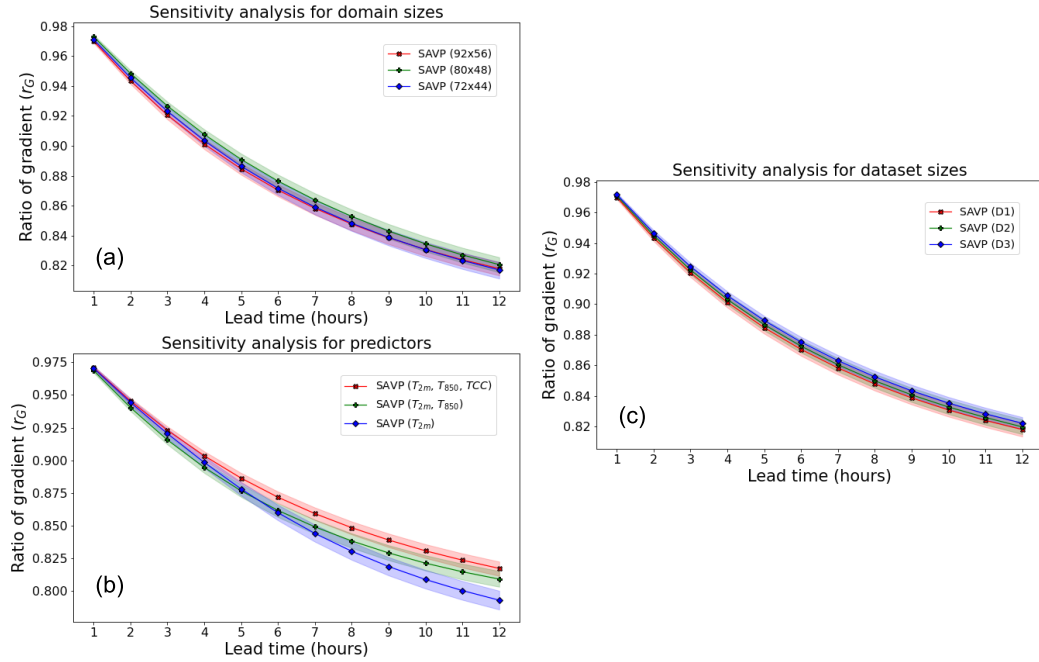



Figure A5. Like Fig. A3, but for the gradient ratio r_G.

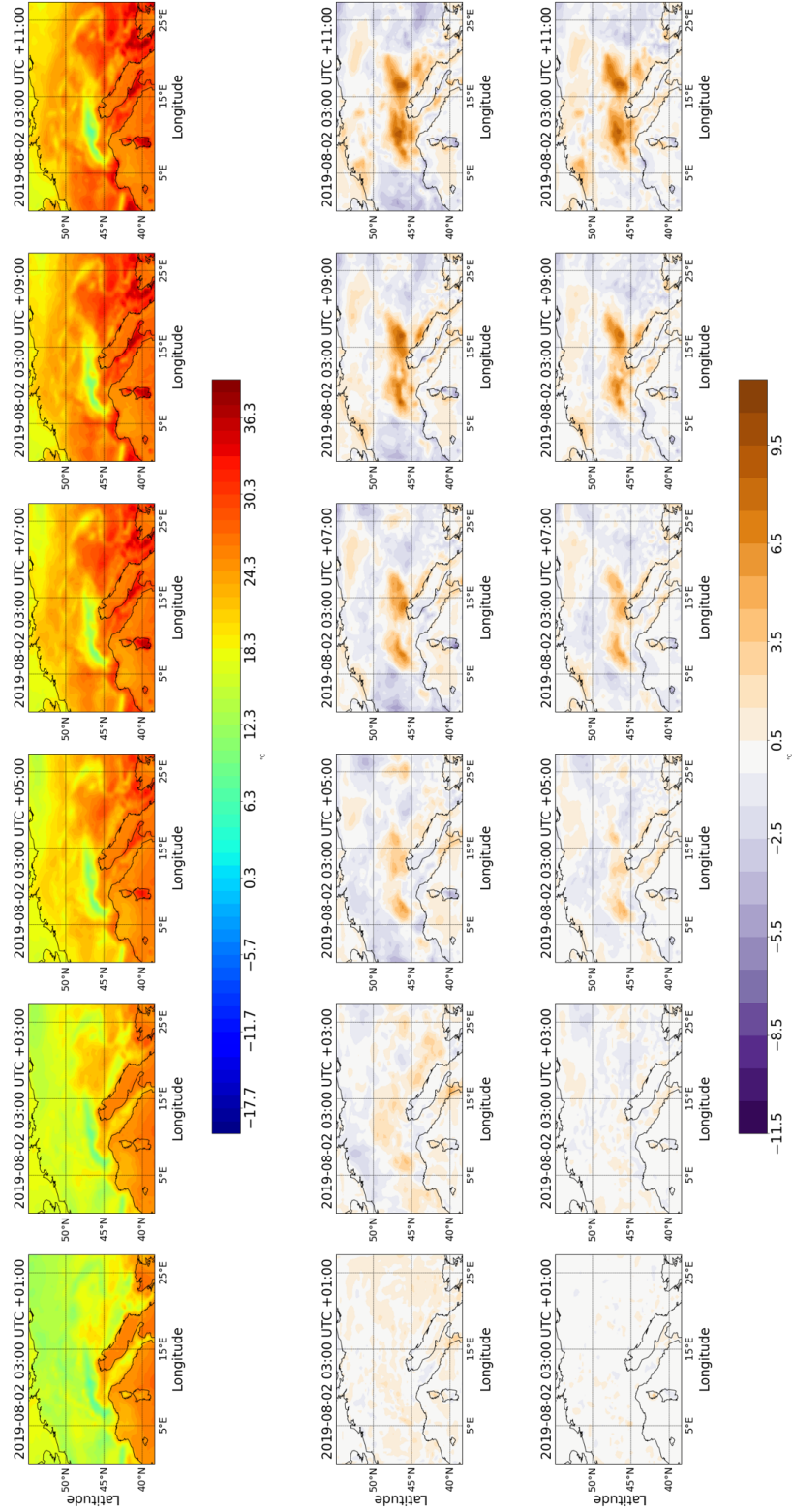



Figure A6. Analogous to Fig. 6, but for additional lead times (1, 3, 5, 7, 9 and 11 hours).

Author contributions. Author contributions. Bing Gong and Michael Langguth equally contributed to this work and performed the bulk of the coding, the experiments, the analysis, and the writing. The study was conceived by Martin Schultz and Bing Gong. Bing Gong, Michael Langguth, Yan Ji, contributed to the method development and maintain the codes. All authors have reviewed and edited the manuscript in several iterations. Martin Schultz supervised the entire project and secured funding.

Competing interests. The authors declare that they have no conflict of interest

Acknowledgements. The authors acknowledge funding from the DeepRain project under grant agreement 01 IS18047A from the Bundesministerium für Bildung und Forschung (BMBF), from the European Union H2020 MAELSTROM project (grant No. 955513, co-funding by BMBF), and from the ERC Advanced grant IntelliAQ (grant no. 787576). We thank Olaf Stein and Lars Hoffmann for preparing the datasets used in our research, as well as Severin Hußmann for an initial attempt to apply video prediction techniques to weather forecasting and helpful scientific discussions. The authors also gratefully acknowledge the Helmholtz Association's Earth System Modelling Project (ESM) for funding this work by providing computing time on the ESM partition of the supercomputer JUWELS as it described by Jülich Supercomputing Centre (2019) at the Jülich Supercomputing Centre (JSC).

References

Ayzel, G., Heistermann, M., and Winterrath, T.: Optical flow models as an open benchmark for radar-based precipitation nowcasting (rainy-motion v0. 1), Geoscientific Model Development, 12, 1387–1402, 2019.

680   Ayzel, G., Scheffer, T., and Heistermann, M.: RainNet v1. 0: a convolutional neural network for radar-based precipitation nowcasting, Geoscientific Model Development, 13, 2631–2644, 2020.

Bauer, P., Thorpe, A., and Brunet, G.: The quiet revolution of numerical weather prediction, Nature, 525, 47–55, https://doi.org/https://doi.org/10.1038/nature14956, 2015.

Bihlo, A.: A generative adversarial network approach to (ensemble) weather prediction, arXiv:2006.07718 [physics, stat], 2020.

685   Brenowitz, N. D., Beucler, T., Pritchard, M., and Bretherton, C. S.: Interpreting and stabilizing machine-learning parametrizations of convection, Journal of the Atmospheric Sciences, 77, 4357–4375, 2020.

Brock, A., Donahue, J., and Simonyan, K.: Large Scale GAN Training for High Fidelity Natural Image Synthesis, in: 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019, OpenReview.net, https://openreview.net/forum?id=B1xsqj09Fm, 2019.

690   Canziani, A., Paszke, A., and Culurciello, E.: An analysis of deep neural network models for practical applications, arXiv preprint arXiv:1605.07678, 2016.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A.: Emerging Properties in Self-Supervised Vision Transformers, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 9650–9660, 2021.

Casaioli, M., Mantovani, R., Proietti Scorzoni, F., Puca, S., Speranza, A., and Tirozzi, B.: Linear and nonlinear post-processing of numerically
695   forecasted surface temperature, Nonlinear Processes in Geophysics, 10, 373–383, https://doi.org/10.5194/npg-10-373-2003, 2003.

Chantry, M., Hatfield, S., Dueben, P., Polichtchouk, I., and Palmer, T.: Machine learning emulation of gravity wave drag in numerical weather forecasting, Journal of Advances in Modeling Earth Systems, 13, e2021MS002 477, 2021.

Chattopadhyay, A., Nabizadeh, E., and Hassanzadeh, P.: Analog forecasting of extreme-causing weather patterns using deep learning, Journal of Advances in Modeling Earth Systems, 12, e2019MS001 958, 2020.

700   Chepfer, H., Brogniez, H., and Noël, V.: Diurnal variations of cloud and relative humidity profiles across the tropics, Scientific reports, 9, 1–9, 2019.

Clark, A., Donahue, J., and Simonyan, K.: Adversarial video generation on complex datasets, arXiv preprint arXiv:1907.06571, 2019.

de Bézenac, E., Pajot, A., and Gallinari, P.: Deep learning for physical processes: incorporating prior scientific knowledge, Journal of Statistical Mechanics: Theory and Experiment, 2019, 124 009, https://doi.org/10.1088/1742-5468/ab3195, 2019.

705   Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database, in: CVPR09, 2009.

Denton, E. and Fergus, R.: Stochastic video generation with a learned prior, arXiv preprint arXiv:1802.07687, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N.: An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, arXiv preprint arXiv:2010.11929, 2020.

710   Dueben, P. D. and Bauer, P.: Challenges and design choices for global weather and climate models based on machine learning, Geoscientific Model Development, 11, 3999–4009, 2018.

Ebert, F., Finn, C., Lee, A. X., and Levine, S.: Self-supervised visual planning with temporal skip connections, arXiv preprint arXiv:1710.05268, 2017.

Eccel, E., Ghielmi, L., Granitto, P., Barbiero, R., Grazzini, F., and Cesari, D.: Prediction of minimum temperatures in an alpine region by linear and non-linear post-processing of meteorological models, Nonlinear Processes in Geophysics, 14, 211–222, https://doi.org/10.5194/npg-14-211-2007, 2007.

Efron, B. and Tibshirani, R. J.: An introduction to the bootstrap, CRC press, https://doi.org/10.1201/9780429246593, 1994.

Espeholt, L., Agrawal, S., Sønderby, C., Kumar, M., Heek, J., Bromberg, C., Gazen, C., Hickey, J., Bell, A., and Kalchbrenner, N.: Skillful Twelve Hour Precipitation Forecasts using Large Context Neural Networks, arXiv preprint arXiv:2111.07470, 2021.

Franceschi, J.-Y., Delasalles, E., Chen, M., Lamprier, S., and Gallinari, P.: Stochastic latent residual video prediction, in: International Conference on Machine Learning, pp. 3233–3246, PMLR, 2020.

Fuks, O. and Tchelepi, H. A.: Limitations of physics informed machine learning for nonlinear two-phase transport in porous media, Journal of Machine Learning for Modeling and Computing, 1, 2020.

Garratt, J. R.: The atmospheric boundary layer, Earth-Science Reviews, 37, 89–134, 1994.

Ginzburg, A. I., Kostianoy, A. G., and Sheremet, N. A.: Sea surface temperature variability, in: The Black Sea Environment, pp. 255–275, Springer, 2007.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y.: Generative adversarial nets, in: Advances in neural information processing systems, pp. 2672–2680, 2014.

Grönquist, P., Yao, C., Ben-Nun, T., Dryden, N., Dueben, P., Li, S., and Hoefler, T.: Deep learning for post-processing ensemble weather forecasts, Philosophical Transactions of the Royal Society A, 379, 20200 092, 2021.

Guen, V. L. and Thome, N.: Disentangling physical dynamics from unknown factors for unsupervised video prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11 474–11 484, 2020.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A.: Improved Training of Wasserstein GANs, 2017.

Haiden, T., Janousek, M., Vitart, F., Ben-Bouallegue, Z., Ferranti, L., and Prates, F.: Evaluation of ECMWF forecasts, including the 2021 upgrade, https://doi.org/10.21957/90pgicjk4, 2021.

Han, Y., Zhang, G. J., Huang, X., and Wang, Y.: A moist physics parameterization based on deep learning, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 076, 2020.

Hatfield, S., Chantry, M., Dueben, P., Lopez, P., Geer, A., and Palmer, T.: Building Tangent-Linear and Adjoint Models for Data Assimilation With Neural Networks, Journal of Advances in Modeling Earth Systems, 13, e2021MS002 521, 2021.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., et al.: The ERA5 global reanalysis, Quarterly Journal of the Royal Meteorological Society, 146, 1999–2049, 2020.

Huth, R.: Statistical downscaling of daily temperature in central Europe, Journal of Climate, 15, 1731–1742, 2002.

Huth, R.: Sensitivity of local daily temperature change estimates to the selection of downscaling models and predictors, Journal of Climate, 17, 640–652, 2004.

Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A.: Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1125–1134, 2017.

Jin, B., Hu, Y., Tang, Q., Niu, J., Shi, Z., Han, Y., and Li, X.: Exploring spatial-temporal multi-frequency analysis for high-fidelity and temporal-consistency video prediction, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4554–4563, 2020.

Jin, X., Cai, S., Li, H., and Karniadakis, G. E.: NSFnets (Navier-Stokes flow nets): Physics-informed neural networks for the incompressible Navier-Stokes equations, Journal of Computational Physics, 426, 109 951, 2021.

Jülich Supercomputing Centre: JUWELS: Modular Tier-0/1 Supercomputer at the Jülich Supercomputing Centre, Journal of large-scale research facilities, 5, https://doi.org/10.17815/jlsrf-5-171, 2019.

Karniadakis, G. E., Kevrekidis, I. G., Lu, L., Perdikaris, P., Wang, S., and Yang, L.: Physics-informed machine learning, Nature Reviews Physics, 3, 422–440, 2021.

Karpatne, A., Watkins, W., Read, J., and Kumar, V.: Physics-guided neural networks (pgnn): An application in lake temperature modeling, arXiv preprint arXiv:1710.11431, 2017.

Khain, A., Beheng, K., Heymsfield, A., Korolev, A., Krichak, S., Levin, Z., Pinsky, M., Phillips, V., Prabhakaran, T., Teller, A., et al.: Representation of microphysical processes in cloud-resolving models: Spectral (bin) microphysics versus bulk parameterization, Reviews of Geophysics, 53, 247–322, 2015.

Kim, S., Hong, S., Joh, M., and Song, S.-k.: Deeprain: Convlstm network for precipitation prediction using multichannel radar data, arXiv preprint arXiv:1711.02316, 2017.

Kingma, D. P. and Ba, J.: Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980, 2014.

Kingma, D. P. and Welling, M.: Auto-encoding variational bayes, arXiv preprint arXiv:1312.6114, 2013.

Kong, Y. and Fu, Y.: Human action recognition and prediction: A survey, arXiv preprint arXiv:1806.11230, 2018.

Krizhevsky, A., Sutskever, I., and Hinton, G. E.: ImageNet Classification with Deep Convolutional Neural Networks, in: Advances in Neural Information Processing Systems, edited by Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., vol. 25, Curran Associates, Inc., https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf, 2012.

Lee, A. X., Zhang, R., Ebert, F., Abbeel, P., Finn, C., and Levine, S.: Stochastic adversarial video prediction, arXiv preprint arXiv:1804.01523, 2018.

Leinonen, J., Nerini, D., and Berne, A.: Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network, IEEE Transactions on Geoscience and Remote Sensing, https://doi.org/10.1109/TGRS.2020.3032790, 2020.

Lezama Valdes, L.-M., Katurji, M., and Meyer, H.: A Machine Learning Based Downscaling Approach to Produce High Spatio-Temporal Resolution Land Surface Temperature of the Antarctic Dry Valleys from MODIS Data, Remote Sensing, 13, 4673, 2021.

Liu, H.-B. and Lee, I.: MPL-GAN: Toward Realistic Meteorological Predictive Learning Using Conditional GAN, IEEE Access, 8, 93 179–93 186, 2020.

Liu, W., Luo, W., Lian, D., and Gao, S.: Future frame prediction for anomaly detection–a new baseline, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 6536–6545, 2018.

Liu, Y., Key, J. R., and Wang, X.: The influence of changes in cloud cover on recent surface temperature trends in the Arctic, Journal of Climate, 21, 705–715, 2008.

Lorenz, E. N.: The predictability of a flow which possesses many scales of motion, Tellus, 21, 289–307, https://doi.org/10.3402/tellusa.v21i3.10086, 1969.

Mathieu, M., Couprie, C., and LeCun, Y.: Deep multi-scale video prediction beyond mean square error, arXiv preprint arXiv:1511.05440, 2015.

McGovern, A., Elmore, K. L., Gagne, D. J., Haupt, S. E., Karstens, C. D., Lagerquist, R., Smith, T., and Williams, J. K.: Using artificial intelligence to improve real-time decision-making for high-impact weather, Bulletin of the American Meteorological Society, 98, 2073–2090, 2017.

Murphy, A. H. and Winkler, R. L.: A general framework for forecast verification, Monthly weather review, 115, 1330–1338, 1987.

Oliu, M., Selva, J., and Escalera, S.: Folded recurrent neural networks for future video prediction, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 716–731, 2018.

Oprea, S., Martinez-Gonzalez, P., Garcia-Garcia, A., Castro-Vargas, J. A., Orts-Escolano, S., Garcia-Rodriguez, J., and Argyros, A.: A review on deep learning techniques for video prediction, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020.

Orlanski, I.: A rational subdivision of scales for atmospheric processes, Bulletin of the American Meteorological Society, pp. 527–530, https://doi.org/https://doi.org/10.1175/1520-0477-56.5.527, 1975.

Owens, R. G. and Hewson, T. D.: ECMWF Forecast User Guide, Tech. rep., 2018.

Patraucean, V., Handa, A., and Cipolla, R.: Spatio-temporal video autoencoder with differentiable memory, arXiv preprint arXiv:1511.06309, 2015.

Prudden, R., Adams, S., Kangin, D., Robinson, N., Ravuri, S., Mohamed, S., and Arribas, A.: A review of radar-based nowcasting of precipitation and applicable machine learning techniques, arXiv preprint arXiv:2005.04988, 2020.

Qi, M., Wang, Y., Li, A., and Luo, J.: STC-GAN: Spatio-temporally coupled generative adversarial networks for predictive scene parsing, IEEE Transactions on Image Processing, 29, 5420–5430, 2020.

Racah, E., Beckham, C., Maharaj, T., Prabhat, and Pal, C. J.: Semi-Supervised Detection of Extreme Weather Events in Large Climate Datasets, CoRR, abs/1612.02095, http://arxiv.org/abs/1612.02095, 2016.

Raissi, M., Perdikaris, P., and Karniadakis, G. E.: Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations, Journal of Computational physics, 378, 686–707, 2019.

Raissi, M., Yazdani, A., and Karniadakis, G. E.: Hidden fluid mechanics: Learning velocity and pressure fields from flow visualizations, Science, 367, 1026–1030, 2020.

Rao, C., Sun, H., and Liu, Y.: Physics-informed deep learning for incompressible laminar flows, Theoretical and Applied Mechanics Letters, 10, 207–212, 2020.

Rao, Q. and Frtunikj, J.: Deep learning for self-driving cars: Chances and challenges, in: Proceedings of the 1st International Workshop on Software Engineering for AI in Autonomous Systems, pp. 35–38, 2018.

Rasp, S. and Lerch, S.: Neural networks for postprocessing ensemble weather forecasts, Monthly Weather Review, 146, 3885–3900, https://doi.org/10.1175/MWR-D-18-0187.1, 2018.

Rasp, S. and Thuerey, N.: Data-Driven Medium-Range Weather Prediction With a Resnet Pretrained on Climate Simulations: A New Model for WeatherBench, Journal of Advances in Modeling Earth Systems, 13, https://doi.org/10.1029/2020MS002405, 2021.

Rasp, S., Dueben, P. D., Scher, S., Weyn, J. A., Mouatadid, S., and Thuerey, N.: WeatherBench: a benchmark data set for data-driven weather forecasting, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 203, 2020.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arribas, A., and Mohamed, S.: Skilful precipitation nowcasting using deep generative models of radar, Nature, 597, 672–677, https://doi.org/10.1038/s41586-021-03854-z, 2021.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al.: Deep learning and process understanding for data-driven Earth system science, Nature, 566, 195–204, 2019.

Reyniers, M.: Quantitative precipitation forecasts based on radar observations: Principles, algorithms and operational systems, Institut Royal Météorologique de Belgique Brussel, Belgium, 2008.

Scher, S.: Toward data-driven weather and climate forecasting: Approximating a simple general circulation model with deep learning, Geophysical Research Letters, 45, 12–616, 2018.

Schultz, M. G., Betancourt, C., Gong, B., Kleinert, F., Langguth, M., Leufen, L. H., Mozaffari, A., and Stadtler, S.: Can deep learning beat numerical weather prediction?, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 379, https://doi.org/10.1098/rsta.2020.0097, 2021.

830 Sha, Y., Gagne II, D. J., West, G., and Stull, R.: Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. Part I: Daily maximum and minimum 2-m temperature, Journal of Applied Meteorology and Climatology, 59, 2057–2073, 2020.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c.: Convolutional LSTM network: A machine learning approach for precipitation nowcasting, vol. 28, 2015.

Sønderby, C. K., Espeholt, L., Heek, J., Dehghani, M., Oliver, A., Salimans, T., Agrawal, S., Hickey, J., and Kalchbrenner, N.: Metnet: A
835 neural weather model for precipitation forecasting, arXiv preprint arXiv:2003.12140, 2020.

Sun, B., Groisman, P. Y., Bradley, R. S., and Keimig, F. T.: Temporal changes in the observed relationship between cloud cover and surface air temperature, Journal of Climate, 13, 4341–4357, 2000.

Sun, J., Xue, M., Wilson, J. W., Zawadzki, I., Ballard, S. P., Onvlee-Hooimeyer, J., Joe, P., Barker, D. M., Li, P.-W., Golding, B., et al.: Use of NWP for nowcasting convective precipitation: Recent progress and challenges, Bulletin of the American Meteorological Society, 95,
840 409–426, 2014.

Sun, J., Xie, J., Hu, J.-F., Lin, Z., Lai, J., Zeng, W., and Zheng, W.-S.: Predicting future instance segmentation with contextual pyramid convlstms, in: Proceedings of the 27th acm international conference on multimedia, pp. 2043–2051, 2019.

Tarek, M., Brissette, F. P., and Arsenault, R.: Evaluation of the ERA5 reanalysis as a potential reference dataset for hydrological modelling over North America, Hydrology and Earth System Sciences, 24, 2527–2544, 2020.

845 Vannitsem, S., Bremnes, J. B., Demaeyer, J., Evans, G. R., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., et al.: Statistical Postprocessing for Weather Forecasts: Review, Challenges, and Avenues in a Big Data World, Bulletin of the American Meteorological Society, 102, E681–E699, 2021.

Villegas, R., Yang, J., Hong, S., Lin, X., and Lee, H.: Decomposing motion and content for natural video sequence prediction, arXiv preprint arXiv:1706.08033, 2017.

850 von Engeln, A. and Teixeira, J.: A planetary boundary layer height climatology derived from ECMWF reanalysis data, Journal of Climate, 26, 6575–6590, 2013.

Wang, S., Teng, Y., and Perdikaris, P.: Understanding and mitigating gradient flow pathologies in physics-informed neural networks, SIAM Journal on Scientific Computing, 43, A3055–A3081, 2021a.

Wang, Y., Long, M., Wang, J., Gao, Z., and Philip, S. Y.: Predrnn: Recurrent neural networks for predictive learning using spatiotemporal
855 lstms, in: Advances in Neural Information Processing Systems, pp. 879–888, 2017.

Wang, Y., Jiang, L., Yang, M.-H., Li, L.-J., Long, M., and Fei-Fei, L.: Eidetic 3d lstm: A model for video prediction and beyond, in: International conference on learning representations, 2018.

Wang, Y., Wu, H., Zhang, J., Gao, Z., Wang, J., Yu, P. S., and Long, M.: PredRNN: A Recurrent Neural Network for Spatiotemporal Predictive Learning, arXiv preprint arXiv:2103.09504, 2021b.

860 Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P.: Image quality assessment: from error visibility to structural similarity, IEEE transactions on image processing, 13, 600–612, 2004.

Weyn, J. A., Durran, D. R., and Caruana, R.: Can machines learn to predict weather? Using deep learning to predict gridded 500-hPa geopotential height from historical weather data, Journal of Advances in Modeling Earth Systems, 11, 2680–2693, 2019.

Weyn, J. A., Durran, D. R., and Caruana, R.: Improving data-driven global weather prediction using deep convolutional neural networks on a cubed sphere, Journal of Advances in Modeling Earth Systems, 12, e2020MS002 109, 2020.

Wilks, D. S.: Statistical methods in the atmospheric sciences, vol. 100, Academic press, 2011.

Wilson, J. W., Feng, Y., Chen, M., and Roberts, R. D.: Nowcasting challenges during the Beijing Olympics: Successes, failures, and implications for future nowcasting systems, Weather and Forecasting, 25, 1691–1714, 2010.

Yan, W., Zhang, Y., Abbeel, P., and Srinivas, A.: VideoGPT: Video Generation using VQ-VAE and Transformers, arXiv preprint arXiv:2104.10157, 2021.

Zaengl, G., Reinert, D., Rípodas, P., and Baldauf, M.: The ICON (ICOsahedral Non-hydrostatic) modelling framework of DWD and MPI-M: Description of the non-hydrostatic dynamical core, Quarterly Journal of the Royal Meteorological Society, 141, 563–579, 2015.

Ziyin, L., Hartwig, T., and Ueda, M.: Neural networks fail to learn periodic functions and how to fix it, arXiv preprint arXiv:2006.08195, 2020.

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.