Environmental Data Science (2023), 2: e6, 1–24
doi:10.1017/eds.2023.1

METHODS PAPER

AtmoDist: Self-supervised representation learning for atmospheric dynamics

Sebastian Hoffmann and Christian Lessig*

Institut für Simulation und Graphik, Otto-von-Guericke-Universität Magdeburg, Magdeburg, Germany
*Corresponding author. E-mail: christian.lessig@ovgu.de

Received: 01 February 2022; **Revised:** 23 August 2022; **Accepted:** 12 December 2022

Keywords: Downscaling; pretext task; representation learning; self-supervised learning; super-resolution

Abstract

Representation learning has proven to be a powerful methodology in a wide variety of machine-learning applications. For atmospheric dynamics, however, it has so far not been considered, arguably due to the lack of large-scale, labeled datasets that could be used for training. In this work, we show how to sidestep the difficulty and introduce a self-supervised learning task that is applicable to a wide variety of unlabeled atmospheric datasets. Specifically, we train a neural network on the simple yet intricate task of predicting the temporal distance between atmospheric fields from distinct but nearby times. We demonstrate that training with this task on the ERA5 reanalysis dataset leads to internal representations that capture intrinsic aspects of atmospheric dynamics. For example, when employed as a loss function in other machine-learning applications, the derived AtmoDist distance leads to improved results compared to the ℓ_2-loss. For downscaling one obtains higher resolution fields that match the true statistics more closely than previous approaches and for the interpolation of missing or occluded data the AtmoDist distance leads to results that contain more realistic fine-scale features. Since it is obtained from observational data, AtmoDist also provides a novel perspective on atmospheric predictability.

Impact Statement

This work demonstrates that the tenet of representation learning also applies to atmospheric dynamics and that the intermediate activations of trained neural networks can provide an informative embedding for atmospheric data. We show this with a novel, self-supervised learning task derived from the underlying physics that makes large amounts of unlabeled observational and simulation data accessible for machine learning. We exemplify the utility of our learned representations by deriving from these a distance metric for atmospheric dynamics that leads to improved performance in applications such as downscaling and missing data interpolation.

1. Introduction

Representation learning is an important methodology in machine learning where the focus is on the data transformations that are provided by a neural network. The motivation for it is to obtain an embedding of the input data that will facilitate a range of applications, for example, by revealing intrinsic aspects of the data or by being invariant to perturbations that are irrelevant to tasks. Representation learning is today central to application areas such as machine translation (e.g., Devlin et al., 2019; Brown et al., 2020), and

image understanding (e.g., Caron et al., 2021; Bao et al., 2022; He et al., 2022), and has led there to significantly improved performance on a variety of tasks.

In the geosciences, representation learning has so far received only limited attention. One reason is the lack of large-scale, labeled datasets that are classically used for training. As has been shown for other domains (e.g., Devlin et al., 2019; He et al., 2020; Caron et al., 2021), representation learning can, however, benefit from working with unlabeled data and performing self-supervised learning with loss functions derived from the data itself. One reason for this is that a self-supervised task can be more challenging than, for example, choosing from a small set of possible answers or labels. Hence, with a self-supervised task the neural network is forced to learn more expressive and explanatory internal representations. A second reason for the efficiency of self-supervised training is that it makes much larger amounts of data available since no labels are required (e.g., Devlin et al., 2019; Brown et al., 2020; Zhai et al., 2021).

In this work, we introduce self-supervised representation learning for atmospheric dynamics and demonstrate its utility by defining a novel, data-driven distance metric for atmospheric states based on it. For the self-supervised training, we propose a novel learning task that is applicable to a wide range of datasets in atmospheric science. Specifically, given a temporal sequence of datum, for example, spatial fields in a reanalysis or from a simulation, the task of the neural network is to predict the temporal distance between two randomly selected, close-by sequence elements. Performing well on the task requires the network to develop an internal representation of the underlying dynamics, which will typically be useful for a range of tasks.

We demonstrate the effectiveness and practicality of our self-supervised training task by learning a representation network for vorticity and divergence (which are equivalent to the wind velocity field) from ERA5 reanalysis (Hersbach et al., 2020), see Figure 1 for an overview. From the learned representation, we subsequently derive a data-driven distance metric for atmospheric states, which we call the AtmoDist distance. To demonstrate its potential, we use it as a loss function in the generative adversarial network (GAN)-based downscaling as well as to interpolate missing data. Building on the state-of-the-art GAN by Stengel et al. (2020) we show that the AtmoDist loss significantly improves downscaling results compared to the ℓ_2-loss used in the original work. For missing data interpolation, AtmoDist leads to more realistic fine-scale details and better local statistics. As a baseline, we use in the experiments an auto-encoder, which provides an alternative means to obtain a feature space representation with self-supervised training. We also report results on experiments with AtmoDist on the predictability of atmospheric states where the data-driven loss reproduces known dependencies on season and spatial location.

The tasks evaluated in our work, that is, downscaling and missing data interpolation, are only two possible applications that we selected to demonstrate the usefulness of the learned representations of AtmoDist. Forecasting (Rasp et al., 2020; Bi et al., 2022; Lam et al., 2022; Pathak et al., 2022), climate response modeling (Watson-Parris et al., 2022), or the detection and modeling of extreme weather events (Racah et al., 2017; Blanchard et al., 2022), are other potential tasks that could benefit from improved representation learning.

We believe that self-supervised representation learning for atmospheric data has significant potential and we hence consider the present work as a first step in this direction. Self-supervised learning only requires unlabeled data, which is available in significant quantities, for example, in the form of satellite observations, reanalyses, and simulation outputs. Given the difficulty of obtaining large, labeled datasets, this removes an obstacle to the use of large-scale machine learning in the atmospheric sciences. At the same time, representation learning can "distill" effective representations from very large amounts of data (Devlin et al., 2019; Zhai et al., 2021), which might, for example, provide a new avenue to process the outputs produced by large simulation runs (Eyring et al., 2016). We believe that learned representation can also be useful to gain novel scientific insights into the physics, similar to how proper orthogonal decompositions have been used in the past. This is, in our opinion, a particularly inspiring direction for future work (Toms et al., 2020).

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

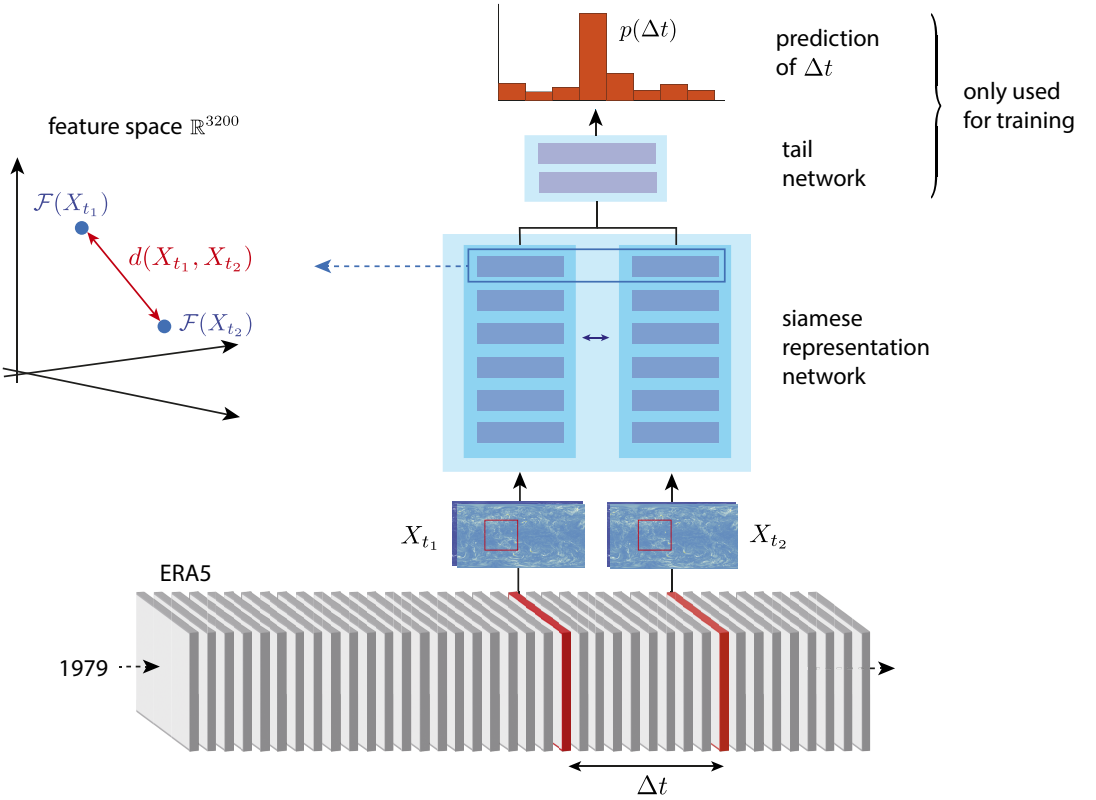

Figure 1. *Overview of the methodology for AtmoDist. From a temporal sequence of atmospheric fields (bottom), two nearby ones are selected at random (red) and stored together with their temporal separation Δt as a training sample. Both fields are then passed through the same representation network (blue), which embeds them into a high-dimensional feature space (left). These embeddings are subsequently used by the tail network to predict the temporal separation Δt (top, orange). The whole architecture is trained end-to-end. Once training is completed, the embeddings can be used in downstream tasks, for example, through a distance measure $d(X_{t_1}, X_{t_2})$ in embedding space.*

2. Related Work

In the following, we will discuss pertinent related work from the geosciences and machine learning.

2.1. Geosciences

Distance measures for atmospheric states play an important role in classical weather and climate predictions. For example, ensemble methods require a well-defined notion of nearby atmospheric states for their initialization. Various distance measures have therefore been proposed in the literature, typically grounded in mathematical and physical considerations, for example, conservation laws. The importance of an appropriate distance measure for atmospheric states already appears in the classical work by Lorenz (1969) where atmospheric predictability depends on the closeness of initial states and is also affected by the characteristics of their spectrum that is, a Sobolev-type measure. Talagrand (1981) considered an energy metric around a reference state obtained from the primitive equations in work on 4D data assimilation. Palmer et al. (1998) argue that within the framework of linearized equations and with singular vectors as coordinates, a metric for targeting observations should not only be informed by geophysical fluid dynamics considerations but also consider the operational

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

observing network. Recently, Koh and Wan (2015) introduce an energy metric that does not require an atmospheric reference state but is intrinsically defined. For the case of an ideal barotropic fluid, the metric of Koh and Wan (2015) also coincides with the geodesic metric that was introduced by Arnold (1966) and studied by Ebin and Marsden (1970) to describe the fluid motion as a geodesic on the infinite-dimensional group of volume preserving diffeomorphisms. Although of utility in classical applications, the aforementioned distance measures lack the sensitivity desirable for machine-learning techniques, for example, with respect to small-scale features, and are agnostic to applications. In the context of downscaling, this deficiency has recently been noted by Stengel et al. (2020).

2.2. Representation learning and learned distance measures

Representation learning (Bengio et al., 2013) focuses on the nonlinear transformations that are realized by a neural network and understands these as a mapping of the input data to a feature space adapted to the data domain. The feature space is informative and explanatory, for example, when different classes are well separated and interdependencies are transparently encoded. This then allows to solve so-called downstream applications in a simple and efficient manner, for example, by appending a linear transformation or a small neural network to the pre-trained one. Good representations will also be useful for a wide range of applications.

A pertinent example of the importance of representations in neural networks is classification. There, the bulk of the overall network architecture is usually devoted to transforming the data into a feature space where the different classes correspond to linear and well-separated subspaces. A linear mapping in the classification head then suffices to accurately solve the task and the entire preceding network can thus be considered as a representation one. With deep neural networks, one obtains a hierarchy of representations where deeper once typically correspond to more abstract features, see, for example, Zeiler and Fergus (2014) for visualizations. The hierarchical structure is of importance, for example, for generative machine-learning models (e.g., Ronneberger et al., 2015; Karras et al., 2019, 2020; Ranftl et al., 2021) where features at all scales have to match the target distribution.

An important applications of representation learning is the design of domain-specific loss functions, sometimes also denoted as content losses (Zhang et al., 2018). The rationale for these is that feature spaces are designed to capture the essential aspects of an input data domain and computing a distance there is hence more discriminative than on the raw inputs (Achille and Soatto, 2018). Furthermore, deeper layers typically have invariance against "irrelevant" perturbations, for example, translation, rotation, and noise in images. A domain where content losses play an important role is computer vision where ℓ_p-norms in the pixel domain of images are usually not well suited for machine learning, for example, because a small shift in the image content can lead to a large distance in an ℓ_p-norm despite the image being semantically unchanged. Loss functions computed in the feature spaces of networks such as VGG (Simonyan and Zisserman, 2015), in contrast, can lead to substantially improved performance in task such as in-painting (Yang et al., 2017), style transfer (Gatys et al., 2016), and image synthesis (Ledig et al., 2017; Karras et al., 2019).

2.3. Self-supervised learning

Closely related to representation learning is self-supervised learning that is today the state-of-the-art methodology for obtaining informative and explanatory representations. The appeal of self-supervised learning is that it does not require labeled data but uses for training a loss function that solely depends on the data itself. In computer vision, for example, a common self-supervised learning task is to in-paint (or predict) a region that was cropped out from a given image (Pathak et al., 2016). Since training is typically informed by the data and not a specific application, self-supervised learning fits naturally with representation learning where one seeks domain- or data-specific but task-independent representations. The ability to use very large amounts of training data, which is usually much easier than in supervised

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

training since no labels are required, also helps in most instances to significantly improve representations (Radford et al., 2018; Devlin et al., 2019; Brown et al., 2020; Zhai et al., 2021).

Prominent examples of pretext tasks for computer vision include solving jigsaw puzzles (Noroozi and Favaro, 2016), learning image rotations (Gidaris et al., 2018), predicting color channels from grayscale images and vice versa (Zhang et al., 2017), or inpainting cropped out regions of an image (Pathak et al., 2016). An early approach that has been used for representation learning is the denoising autoencoder (Vincent et al., 2010). The work of Misra et al. (2016) is directly related to ours in the sense that they train a network to predict the temporal order of a video sequence using a triplet loss. In contrast, our approach relies on predicting the exact (categorical) temporal distance between two patches and not the order, which we believe forces the network to learn more informative representations.

Recently, consistency-based methods have received considerable attention in the literature on self-supervised learning, for example, in the form of contrastive loss functions or student-teacher methods. Since our work employs a pretext task, we will not discuss these methods but refer to Le-Khac et al. (2020) for an overview.

2.4. Machine learning for the geoscience

Deep neural networks have become an important tool in the geosciences in the last years and will likely be relevant for a wide range of problems in the future (e.g., Dueben and Bauer, 2018; Toms et al., 2020; Schultz et al., 2021; Balaji et al., 2022). Reichstein et al. (2019) pointed out the importance of spatiotemporal approaches to machine learning in the field. This implies the need for network architectures adapted to spatiotemporal data as well as suitable learning protocols and loss functions. AtmoDist addresses the last aspect.

An early example of spatial representation learning in the geosciences is Tile2Vec (Jean et al., 2019) for remote sensing. The work demonstrates that local tiles, or patches, can serve as analogues to words in geospatial data and that this allows for the adoption of ideas from natural language processing to geoscience. Related is Space2Vec which can be considered as a multi-scale representation learning approach of geolocations. A spatiotemporal machine-learning approach is used by Barnes et al. (2018) where a neural network is trained to predict the global year of a temperature field. This is similar to AtmoDist where the training task is also the prediction of temporal information given a spatial field. Barnes et al. (2018), however, use their trained network for questions related to global warming whereas we are interested in representation learning. Semi-supervised learning, that is, the combination of a supervised and self-supervised loss function, is employed by Racah et al. (2017) to improve the detection accuracy of extra-tropical and tropical cyclones. As a self-supervised loss, the ℓ_2 reconstruction error of atmospheric fields using an autoencoder is chosen. We also employ an autoencoder as a baseline in this work and find that it leads to significantly worse results on the two downstream tasks we consider.

Apart from the above examples, spatiotemporal representation learning has, to our knowledge, not received widespread attention in atmospheric science. This is in contrast to, for instance, natural language processing (Vaswani et al., 2017; Devlin et al., 2019) or computer vision (Dosovitskiy et al., 2020; Caron et al., 2021; He et al., 2022). In the past, dimensionality reduction techniques such as Principal Component Analysis (PCA), kernel-PCA, or Maximum Variance Unfolding, have been used extensively to analyze atmospheric dynamics (e.g., Lima et al., 2009; Mercer and Richman, 2012; Hannachi and Iqbal, 2019). These can be seen as simple forms of representation learning since they also provide a data transformation to a coordinate system adapted to the input. The need for more expressive representations than those obtained by these methods has been one of the main motivations behind deep neural networks (cf. Bengio et al., 2013).

Several GANs dedicated to geoscience applications have been proposed in the literature (e.g., Stengel et al., 2020; Zhu et al., 2020; Klemmer and Neill, 2021; Klemmer et al., 2022). Noteworthy is SPATE-GAN (Klemmer et al., 2022) that uses a custom metric for spatiotemporal autocorrelation and determines an embedding loss based on it. The authors demonstrate that this improves GAN performance across different datasets without changes to the neural network architecture of the generative model. It hence provides an alternative to AtmoDist proposed in our work. We compare with Stengel et al. (2020) which

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

uses SRGAN (Ledig et al., 2017) to address applications such as wind farm placement. SRGAN was the first work that advocated the use of perceptual loss functions for single-image super-resolution in computer vision. While numerous extensions to the original SRGAN model exist (Lim et al., 2017; Wang et al., 2018), they achieve increased performance by introducing additional complexities to the neural network architecture and training procedure compared to the vanilla SRGAN. In our opinion, this makes them less suited to explore the quality of learned representations and derived loss-functions due to an increased amount of possibly confounding factors. Furthermore, Kurinchi-Vendhan et al. (2021) find that neither of these methods, when applied to atmospheric data, gives a definite advantage over the others and they instead vary in performance depending on what kind of evaluation criterion is used. CEDGAN by Zhu et al. (2020) is an encoder-decoder conditional GAN that is used by the authors for spatial interpolation of orography, a task closely related to super-resolution. The authors solely rely on the adversarial loss term conditioned on the known parts of the field to ensure that the interpolated field is consistent. Such an approach could be easily extended to and possibly benefit from the use of a content-loss that compares the output directly with the ground truth image. GAN-based super-resolution methods could potentially also benefit from additionally conditioning the discriminator on the low-resolution input image. Indeed, the potential to directly use features learned by a discriminator (or generator) has been recognized before (Radford et al., 2015).

3. Method

We perform self-supervised representation learning for atmospheric dynamics and derive a data-driven distance function for atmospheric states from it. For this, we employ a siamese neural network (Chicco, 2021) and combine it with a novel, domain-specific spatiotemporal pretext task that derives from the theory of geophysical fluid dynamics. Specifically, for a given temporal sequence of unlabeled atmospheric states, a neural network is trained to predict the temporal separation between two nearby ones. For the predictions to be accurate, the network has to learn an internal representation that captured the intrinsic properties of atmospheric flows and hence provides feature spaces adapted to atmospheric dynamics. For training, we employ ERA5 reanalysis (Hersbach et al., 2020), which we consider a good approximation to observations. An overview of the AtmoDist methodology is provided in Figure 1.

3.1. Dataset and preprocessing

We employ relative vorticity and divergence to represent an atmospheric state. The two scalar fields are equivalent to the wind velocity vector field, which is the most important dynamic variable and hence a good proxy for the overall state. Our data is from model level 120 of ERA5, which corresponds approximately to the pressure level $883\text{hPa} \pm 85$, and a temporal resolution of 3 hr is used. Vorticity and divergence fields are obtained from the native spectral coefficients of ERA5 by mapping them onto a Gaussian grid with resolution 1280×2560 (we use pyshtools for this [Wieczorek and Meschede, 2018]). The grids are subsequently sampled into patches of size 160×160, which corresponds approximately to $2500\text{km} \times 2500\text{km}$, with randomly selected centers. Following Stengel et al. (2020), we restrict the centers to $\pm 60°$ latitude to avoid the severe distortions close to the poles.

We found that both vorticity and divergence roughly follow a zero-centered Laplace distribution. This led to instabilities, in particular in the training of the downstream task. While clipping values larger than 70 standard deviations was sufficient to stabilize training, this discards information about extreme events that are of particular relevance in many applications. We therefore apply an invertible log-transform to the input data in a preprocessing step and train and evaluate in the log-transformed space, see Appendix A.1 for details.

Training is performed on 3-hourly data from 1979 to 1998 (20 years) while the period from 2000 to 2005 is reserved for evaluation (6 years). This results in $58440 \times N_p$ spatial fields for the training and $17536 \times N_p$ fields for the evaluation set, where N_p is the number of patches per global field of size 1280×2560. We used $N_p = 31$ in our experiments. The maximum time lag, that is, the maximum

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

Table 1. Overview of the data used in this work.

Task	AtmoDist	Super-resolution
Dataset	ERA5	ERA5
Variables	Divergence, Vorticity	Divergence, Vorticity
Model level	120	120
Training period	1979–1998	1979–1998
Evaluation period	2000–2005	2000–2005
Preprocessing	log-space	log-space
Patch-size	160 × 160	96 × 96
Patches per time step	31	180
Center between	60°N–60°S	60°N–60°S
Maximum latitude	82.5°N/82.5°S	73.5°N/73.5°S
Size (training)	741GB	775GB

temporal separation between spatial fields, was $\Delta t_{\max} = 69\,\text{hr}$. This is equivalent to 23 categories for the training of the representation network. An overview of the dataset is given in Table 1.

3.2. Pretext task

Our pretext task is defined for a temporal sequence of spatial fields, for example, atmospheric states from reanalysis or a simulation, and it defines a categorial loss function for self-supervised training. The task is derived from the theory of geophysical fluid dynamics and motivated by the fact that the time evolution of an ideal barotropic fluid is described by a geodesic flow (Arnold, 1966; Ebin and Marsden, 1970). Since a geodesic flow is one of shortest distance, the temporal separation between two nearby states corresponds to an intrinsic distance between them. As a spatiotemporal pretext task for learning a distance measure for atmospheric dynamics, we thus use the prediction of the temporal separation between close-by states. More specifically, given two local patches of atmospheric states X_{t_1}, X_{t_2} centered at the same spatial location but at different, nearby times t_1 and t_2, the task for the neural network is to predict their temporal separation $\Delta t = t_2 - t_1 = n \cdot h_t$ given by a multiple of the time step h_t (3 hr in our case). The categorical label of a tuple (X_{t_1}, X_{t_2}) of input patches, each consisting of the vorticity and divergence field at the respective time $t_k = k \cdot h_t$ for the patch region, is thus defined as the number of time steps n in between them. Following standard methodology for classification problems, for each training item (X_{t_1}, X_{t_2}), our representation network predicts a probability distribution over the finite set of allowed values for n. Training can thus be performed with cross-entropy loss, which is known to be highly effective.

For a distance metric one expects $F(X_{t_1}, X_{t_2}) = F(X_{t_2}, X_{t_1})$. However, we found that reversing the order of inputs results in prediction errors being reversed as well and training the network on randomly ordered pairs did not prevent this behavior. As a consequence, we train the network using a fixed order, that is, we only evaluate $F(X_{t_1}, X_{t_2})$ with $t_1 < t_2$.

3.3. Neural network architecture

The architecture of the neural network we use for representation learning consists of two parts and is schematically depicted in Figure 2. The first part is the representation network. It provides an encoder that maps an atmospheric field X to its feature space representation $\mathcal{F}(X) \subseteq \mathbb{R}^N$. Since both states X_{t_k} of the tuple (X_{t_1}, X_{t_2}) that form a training item are used separately as input to the encoder, it is a siamese network

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

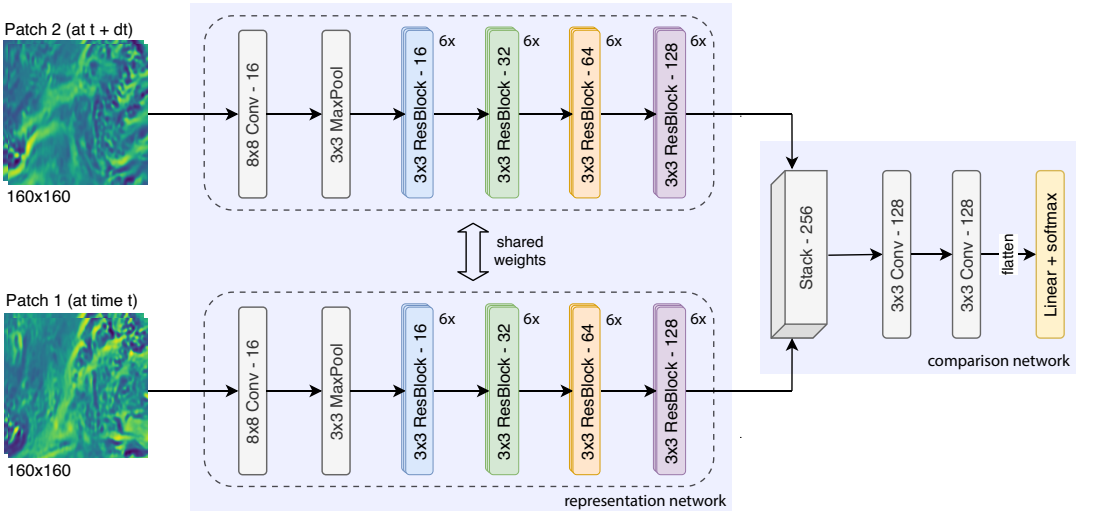

Figure 2. *The AtmoDist network used for learning the pretext task. Numbers after layer names indicate the number of filters/feature maps of an operation. The comparison network is only required during training and can be discarded afterward.*

(Chicco, 2021). The second part of our overall architecture is a tail or comparison network $T(\mathcal{F}(X_{t_1}),\mathcal{F}(X_{t_2}))$ that maps the tuple $(\mathcal{F}(X_{t_1}),\mathcal{F}(X_{t_2}))$ of representations to a probability density $p(\Delta t|X_{t_1},X_{t_2})$ for their temporal separation $\Delta t = n \cdot h_t$. The representation and tail networks are trained simultaneously in an end-to-end manner. After training, only the representation network is of relevance since its activations at the final layer provide the feature space $\mathcal{F}(X)$ for the input X_t that defines the learned representation. The use of activations at intermediate layers is also possible but was not considered in the present work. Note that the tail network should be much smaller than the representation network to facilitate discriminative and explanatory representations.

The representation network follows a residual architecture (He et al., 2015) although with a slightly reduced number of feature maps compared to the standard configurations used in computer vision. It maps an input X of size $2 \times 160 \times 160$ to a representation vector $\mathcal{F}(X)$ of size $128 \times 5 \times 5$. This corresponds to a compression rate of 16. The tail network is a simple convolutional network with a softmax layer at the end to obtain a discrete probability distribution. Both networks together consist of 2,747,856 parameters with 2,271,920 in the encoder and 470,144 in the tail network.

3.4. Training

We train AtmoDist on the dataset described in Section 3.1 using stochastic gradient descent. Since training failed to converge in early experiments, we introduced a pre-training scheme where we initially use only about 10% of the data before switching to the full dataset. For further details on the training procedure, we refer to Appendix A.2.

As can be seen in Figure 3, with pre-training the training loss converges well although overfitting sets in from epoch 27 onwards. Our experiments indicate that the overfitting results from using a fixed, pre-computed set of randomly sampled patches and it could likely be alleviated by sampling these dynamically during training. The noise seen in the evaluation loss is a consequence of the different training and evaluation behavior of the batch normalization layers. While there exist methods to address this issue (Ioffe, 2017), we found them insufficient in our case. Instance normalization (Ulyanov et al., 2017) or layer normalization (Ba et al., 2016) are viable alternatives that should be explored in the future.

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

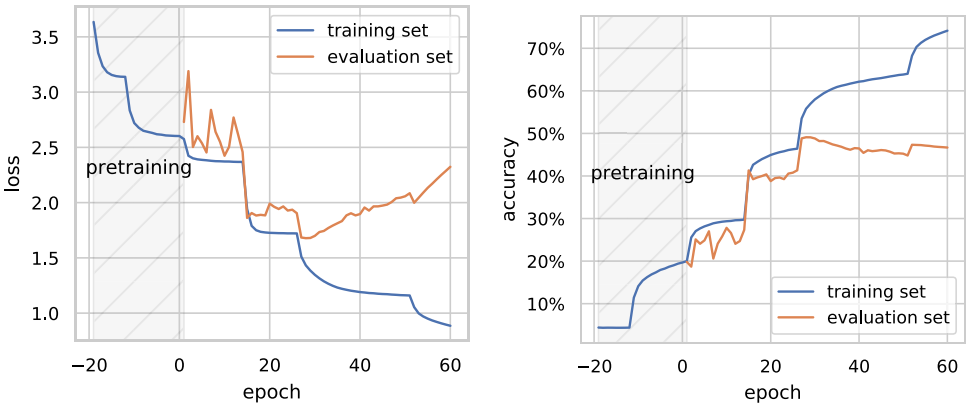

Figure 3. *Loss (left) and Top-1 accuracy (right) during training calculated on the training (1979–1998) and the evaluation dataset (2000–2005). Drops in loss correspond to learning rate reductions. The best loss and accuracy are achieved in epoch 27; afterward the network begins to overfit.*

3.5. Construction of AtmoDist metric

The final layer of the representation network provides an embedding $\mathscr{F}(X_t)$ of the vorticity and divergence fields, which together form X_t, into a feature space (cf. Figure 2). Although this representation can potentially be useful for many different applications, we employ it to define a domain-specific distance functions for atmospheric states.

The feature space representation $\mathscr{F}(X_t)$ is a tensor of size $128 \times 5 \times 5$ that we interpret as a vector, that is, we consider $\mathscr{F}(X_t) \in \mathbb{R}^N$ with $N = 3200$. We then define the AtmoDist metric $d(X_1, X_2)$ for two atmospheric states X_1, X_2 as

$$d_{\text{AtmoDist}}(X_1, X_2) = \frac{1}{N} \|\mathscr{F}(X_1) - \mathscr{F}(X_2)\|^2 \tag{1}$$

where $\|\cdot\|$ denotes the standard ℓ_2-norm. The ℓ_2-norm is commonly used for the construction of metrics based on neural network activations (Gatys et al., 2016; Ledig et al., 2017). Other ℓ_p-norms or weighted norms could potentially also be useful although preliminary experiments indicated that these provide results comparable to Equation (1).

4. Evaluation

The evaluation of representation learning techniques usually employs a collection of downstream applications, since the embedding into the abstract and high-dimensional feature space is in itself rarely insightful. To facilitate interpretation, one thereby typically relies on well-known classification problems. Simple techniques are also employed for the mapping from the representation to the prediction, for example, a small neural network similar to our tail network, to indeed evaluate the representations and not any subsequent computations.

Unfortunately, standardized labeled benchmark datasets akin to MNIST (LeCun et al., 1998) or ImageNet (Russakovsky et al., 2015) currently do not exist for atmospheric dynamics and it is their lack that inspired our self-supervised pretext task. We thus demonstrate the effectiveness of our representations using downscaling, that is, super-resolution, and the interpolation of a partially missing field. Both can be considered as standard problems and have been considered in a variety of previous works (e.g., Requena-Mesa et al., 2019; Groenke et al., 2020; Meraner et al., 2020; Stengel et al., 2020). For downscaling we build on the recent work by Stengel et al. (2020) that provides a state-of-the-art GAN-based downscaling technique and, to facilitate a direct comparison, employ their implementation and replace only the ℓ_2-norm in their code with the AtmoDist distance metric introduced in Section 3.5. For missing data

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

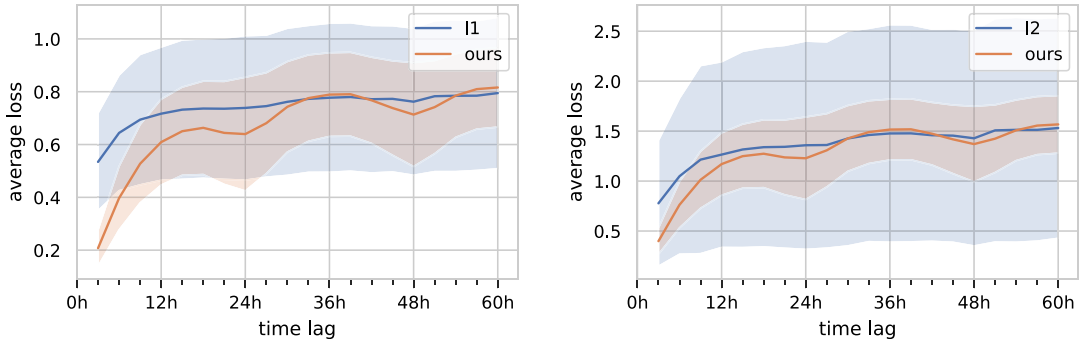

Figure 4. *Mean ℓ_1-norm (left) and mean ℓ_2-norm (right) between samples that are a fixed time-interval apart calculated on the training set. Shaded areas indicate standard deviation. For comparability, the AtmoDist distance has been normalized in each case with the method described in Appendix A.3. To give equal weight to divergence and vorticity, they have been normalized to zero mean and unit variance before calculating grid point-wise metrics.*

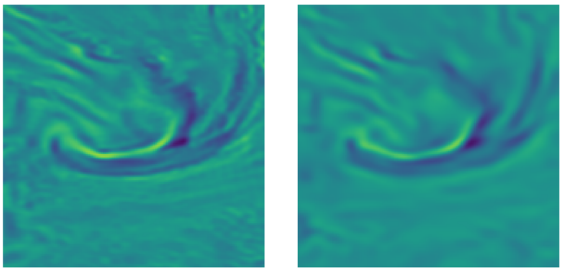

Figure 5. *Divergence field (left) and its reconstruction produced by the autoencoder (right). While the reconstructed field lacks finer details, large-scale structures are properly captured by the autoencoder.*

interpolation, we interpret it as a variant of inpainting and use a network inspired by those successful with the problem in computer vision.

As a baseline, we compare our learned representations against those of an autoencoder (Bengio et al., 2013), one of the earliest representation learning methods. To facilitate a fair comparison, the encoder of the autoencoder is identical to representation network described in Section 3.3. The decoder is a mirrored version of the encoder, replacing downscaling convolutions with upscaling ones. For details, refer to Appendix A.4. After training, the autoencoder is able to produce decent, yet overly -smooth, reconstructions as can be seen in Figure 5.

Before we turn to the downstream applications, we begin with an intrinsic evaluation of the AtmoDist metric using the average distance between atmospheric states with a fixed temporal separation Δt. Since this is close to the training task for AtmoDist, it provides a favorable setting for it. Nonetheless, we believe that the comparison still provides useful insights into our work.

4.1. Intrinsic evaluation of the AtmoDist distance

In order to obtain an intrinsic, application-independent evaluation of the AtmoDist distance in Equation (1), we determine it as a function of temporal separation Δt between two atmospheric states X_{t_1} and X_{t_2}. Note that although the training also employed Δt, the AtmoDist distance metric does no longer use the tail network and the computations are thus different than those during training. Because of the quasi-chaotic nature of the atmosphere (Lorenz, 1969), one expects that any distance measure for it will saturate when the decorrelation time has been reached. To be effective, for example, for machine-learning

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

applications, the distance between states should, however, dependent approximately linear on their temporal separation before the decorrelation time, at least in a statistical sense when a large number of pairs X_{t_1} and X_{t_2} for fixed Δt is considered.

4.1.1. Comparison to ℓ_p-norm

We compute ℓ_1-norm, ℓ_2-norm, and AtmoDist distance as a function of Δt for all atmospheric states that form the training set for AtmoDist and report averaged distances for the different Δt. As shown in Figure 4, the AtmoDist distance takes longer to saturate than mean ℓ_1-norm and ℓ_2-norm and increases more linearly. Also, its standard deviation is significantly smaller and AtmoDist hence provides more consistent distances. Qualitatively similar results are obtained for SSIM (Wang et al., 2004) and PSNR, two popular metric in computer vision, and we report the results for these in Figure 15 in the Appendix.

4.1.2. Temporal behavior

To obtain further insight into the temporal behavior of AtmoDist, we consider the confusion matrix as a function of temporal separation Δt when AtmoDist is used in the training of the network, that is, with the tail network. Figure 6 confirms the expected behavior that predictions get less certain as Δt increases and the states become less correlated. Interestingly, the emergence of sub-diagonals indicates that the network is able to infer the time of the day, that is, the phase of Earth's rotation, with high precision, but it can for large Δt no longer separate different days.

4.1.3. Spatial behavior

The predictability of atmospheric dynamics is not spatially and temporally homogeneous but has a strong dependence on the location as well as the season. One hence would expect that also the error of AtmoDist reflects these intrinsic atmospheric properties. In Figure 7 we show the spatial distribution of the error of AtmoDist, again in the setup used during training with the tail network. As can been seen there, AtmoDist yields good predictions when evaluated near land but performance degrades drastically over the oceans. Apparent in Figure 7 is also a strong difference in predictability between the cold and warm season. This indicates that the model primarily focusses on detecting mesoscale convective activities and not on tracing Lagrangian coherent structures.

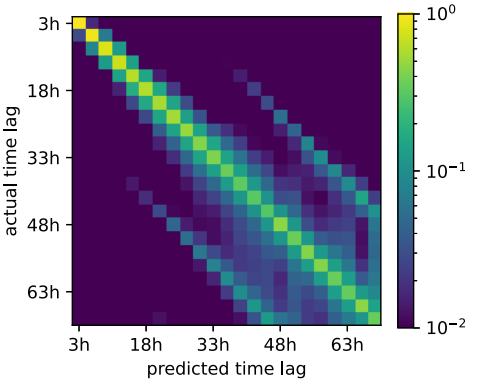

Figure 6. *The confusion matrix shows the accuracy for the evaluation set as a function of predicted time lag and actual time lag. The side-diagonals indicate that AtmoDist is able to infer the time of the day for an atmospheric state with high precision solely based on a local patch of divergence and vorticity fields but might err on the day. A logarithmic color scale has been used to better highlight the side-diagonals.*

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

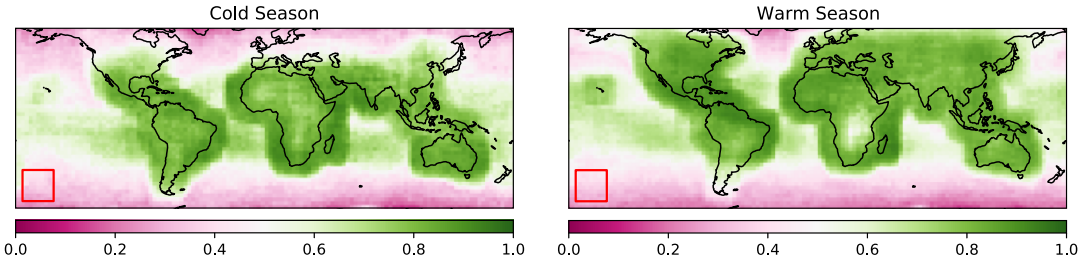

Figure 7. *Accuracy of AtmoDist to correctly predict that two patches are 48 hr apart as a function of space with an error margin of 3 hr (i.e., 45 and 51 hr are also counted as correct prediction). The red rectangle in the lower left corner indicates the patch size used as input for the network.*

4.2. Downscaling

Downscaling, or super-resolution, is a classical problem in both climate science and computer vision. The objective is to obtain a high-resolution field X^{hr} given only a low-resolution version X^{lr} of it. This problem is inherently ill-posed since a given X^{lr} is compatible with a large number of valid high-resolution X^{hr}. Despite this, state-of-the-art methods can often provide valid X^{hr} whose statistics match those of the true fields. In the last years, in particular approaches based on GANs (Goodfellow et al., 2014) have become the de facto standard (e.g., Jiang et al., 2020; Stengel et al., 2020; Klemmer et al., 2022).

Stengel et al. (2020) recently applied GAN-based super-resolution to wind and solar data in North America, demonstrating physically consistent results that outperform competing methods. The authors build on the SRGAN from Ledig et al. (2017) but instead of the VGG network (Simonyan and Zisserman, 2015) that was used as a representation-based content loss in the original work, Stengel et al. (2020) had to resort to an ℓ_2-loss since no analogue for the atmosphere was available. Our work fills this gap and we demonstrate that the learned AtmoDist metric in Equation (1) leads to significantly improved results for atmospheric downscaling. The only modifications to the implementation from Stengel et al. (2020) are a restriction to 4X super-resolution in our work (mainly due to the high computational costs for GAN training), incorporation of an improved initialization scheme for upscaling sub-pixel convolutions (Aitken et al., 2017), as well as replacing transposed convolutions in the generator with regular ones as in the original SRGAN. We also do not use batch normalization in the generator, as suggested by Stengel et al. (2020). For both the ℓ_2-based downscaling as well as the AtmoDist-based downscaling, the model is trained for 18 epochs.

Downscaled divergence fields are shown in Figure 8. Examples for vorticity can be found in Figure 16 in the appendix. Qualitatively, the fields obtained with the AtmoDist metric look sharper than those with an ℓ_2-loss. This overly smooth appearance with ℓ_2-loss is a well-known problem and one of the original motivations for learned content loss functions (Ledig et al., 2017). In Figure 9 (left) we show the average energy spectrum of the downscaled fields. Also with respect to this measure, the AtmoDist metric provides significantly improved results and yields a spectrum very close to the ERA5 ground truth. Following Stengel et al. (2020), we also compare the semivariogram of the downscaled fields that measures the spatial variance of a field $f(x)$ as a function of the lag distance r (Matheron, 1963) (see Appendix A.6 for details on the calculation of the semivariogram). As can be seen in Figure 9 (right) we find that our approach again captures the real geostatistics much better than an ℓ_2-based downscaling. The fields obtained from the autoencoder-based loss are visually of lower quality than with the other two loss functions and the same holds true for the quantitative evaluation metrics (Figure 10).

Finally, we investigate local statistics for the GAN-based downscaling. In Figure 13 (left) we show these for vorticity. The AtmoDist metric again improves the obtained results although a discrepancy to the ERA5 ground truth is still apparent. In Table 2 we report better/worse scores for AtmoDist-based downscaling and those using the ℓ_2-loss for the Wasserstein-1 distance calculated on the empirical distributions (akin to those in Figure 13) for 150 randomly selected, quasi-uniformly distributed cities. A

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

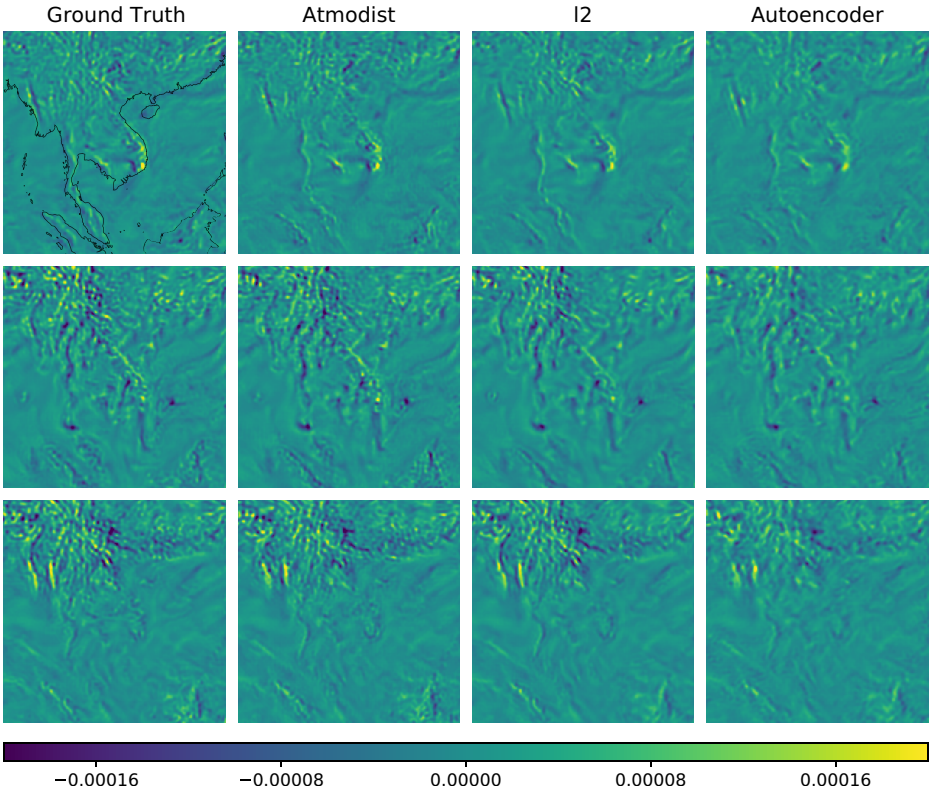

Figure 8. *Uncurated set of downscaled divergence fields over the Gulf of Thailand at different time steps. Coastlines are shown in the first ground truth field and then omitted for better comparability.*

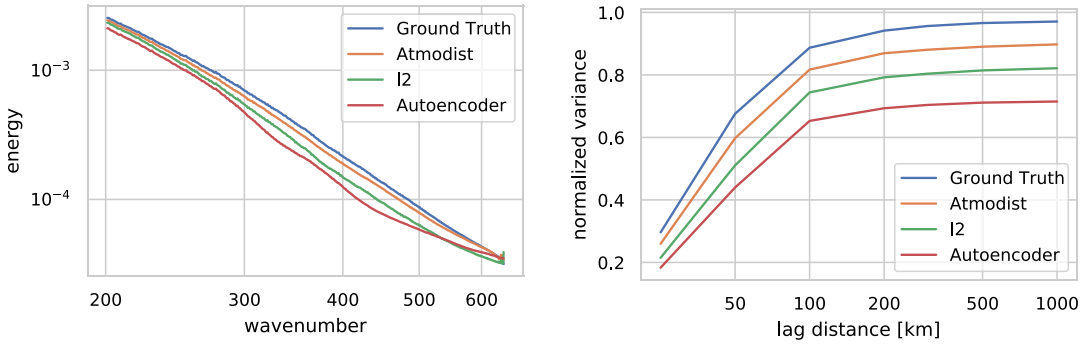

Figure 9. *Left: The energy spectrum starting from wavenumber 200 and averaged over the whole evaluation period. The spectra below wavenumber 200 are almost identical. The spectrum has been calculated by first converting divergence and vorticity to eastwardly and northwardly wind fields, and then evaluating the kinetic energy. Right: Semivariogram of divergence.*

location is thereby scored as better if the Wasserstein-1 distance of the ℓ_2-based super-resolution exceeds 10% of the Wasserstein-1 distance of our approach, and as worse in the opposite case. If neither is the case, that is, both approaches have a comparable error, the location is scored as equal. We find that for divergence we achieve better Wasserstein-1 distances in 102 out of 150 locations while only being worse in 36 out of 150. Similar results are obtained for vorticity.

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

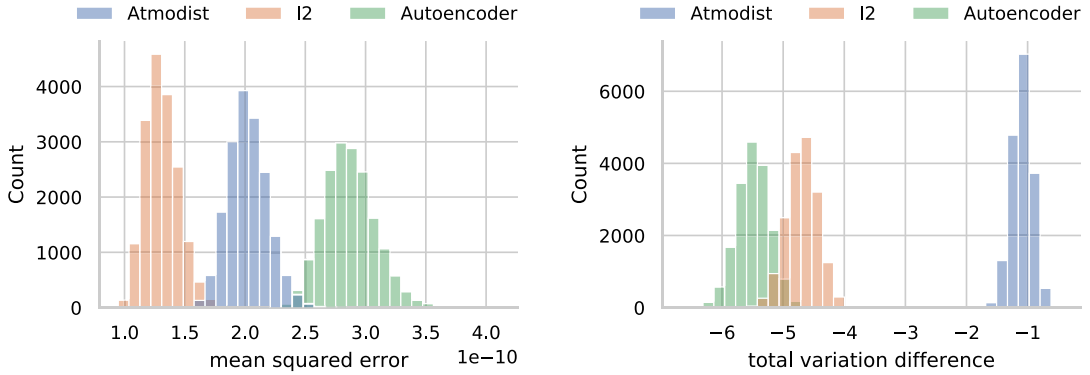

Figure 10. *Histogram of reconstruction errors measured in ℓ_2 norm (left) and difference of total variation (right) for relative vorticity. We define the difference of total variation between the original field f and its super-resolved approximation g as $d_{\text{tv}}(f,g) = \int_{\mathcal{D}} |\nabla f(x)| - |\nabla g(x)| dx$. Values closer to zero are better. Despite performing better with regards to the ℓ_2 reconstruction error, the ℓ_2-based super-resolution performs worse with regards to the difference of total variation. Notice that the approach by Stengel et al. (2020) minimizes the ℓ_2 reconstruction error during training. Interestingly, all three approaches have solely negative total variation differences, implying that the super-resolved fields are overly smooth compared to the ground truth fields. Similar results are obtained for divergence.*

Table 2. Better/worse scores for local statistics of GAN-based super-resolution.

Variable	Better	Equal	Worse
Divergence	102	12	36
Vorticity	90	11	49

4.2.1. Biennial oscillations

In Figure 13 (right) we show the downscaling error for divergence over the six year evaluation period. Clearly visible is an oscillation in the error with a period of approximately two years, which exist also for vorticity and when ℓ_2-loss is used. It is likely that these oscillations are related to the quasi-biennial oscillation (QBO) (Baldwin et al., 2001) and thus reflect intrinsic changes in the predicability in the atmosphere. We leave a further investigation of the effect of the QBO on AtmoDist to future work.

4.3. Reconstruction of partially occluded fields

In the atmospheric sciences, the complete reconstruction of partially occluded or missing fields, for example, because of clouds, is an important problem (e.g., Meraner et al., 2020). It appears in a similar form in computer vision as inpainting.

To further evaluate the performance of AtmoDist, we also use it as a loss for this problem and compare it again against the ℓ_2-loss. For simplicity, we use again ERA5 divergence and vorticity fields as dataset and artificially add occlusion by cropping out a 40×40 region centrally from the 160×160 patches used in the training of the representation network.

As the network for the inpaiting, we choose the identical architecture as for the autoencoder (see Appendix A.4). The 160×160 image with a cropped-out center is passed as input and the network outputs an equally sized image. From that output, only the central region is considered when calculating the loss during training. Details are presented in Appendix A.5.

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

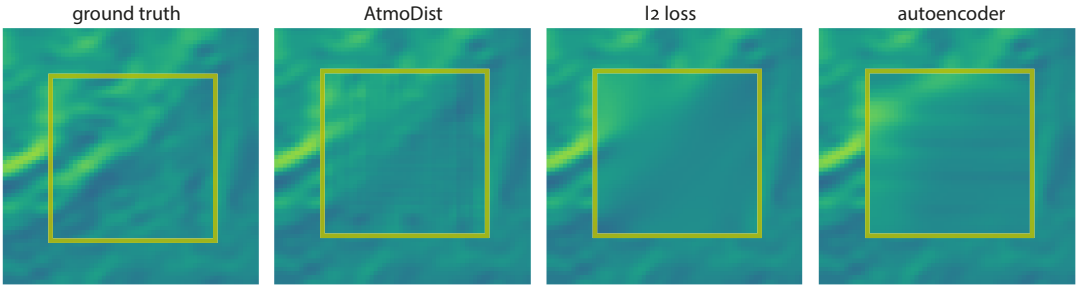

Figure 11. *Interpolated vorticity fields (center, in the yellow square) for deleted regions for, from left to right, ground truth, AtmoDist, ℓ_2-norm, and autoencoder. Both ℓ_2-loss and autoencoder-loss produce overly smooth results. The AtmoDist-based reconstruction captures more of the higher-frequency features present in the data although it suffers from some blocking artifacts. The horizontal artifacts for the autoencoder occurred frequently in our results.*

Figure 11 shows reconstructed fields for all three loss functions. The AtmoDist-based reconstruction produces the most detailed field although it suffers from some blocking artifacts. The ℓ_2-based reconstruction is overly smooth and has homogenous structures instead of fine details. The autoencoder-based field is even smoother than the reconstruction based on the ℓ_2 loss. In fact, we had to tune our training procedure by switching to Adam and lowering the learning rate to prevent this model from generating constant fields. This was not necessary for either the ℓ_2-based or AtmoDist-based reconstructions. Semiovariograms for the reconstructions are shown in Figure 12. These also verify that AtmoDist provides reconstructions with more realistic fine-scale details. The average ℓ_2 norms for AtmoDist, ℓ_2-loss, and autoencoder loss are 0.62, 0.50, and 0.88. Again a clearly better performance of AtmoDist compared to the autoencoder can be observed.

4.4. Ablation study

We performed an ablation study to better understand the effect of the maximum temporal separation $\Delta t_{\max}$ on the performance of AtmoDist. If $\Delta t_{\max}$ is chosen too small, the pretext task might become too easy and a low training error might be achieved with sub-optimal representations. If $\Delta t_{\max}$ is chosen too large, the task might, however, become too difficult and also lead to representations that do not capture the desired effects. We thus trained AtmoDist with $\Delta t_{\max} = \{45\,\text{hr}, 69\,\text{hr}, 93\,\text{hr}\}$ on a reduced dataset with only 66% of

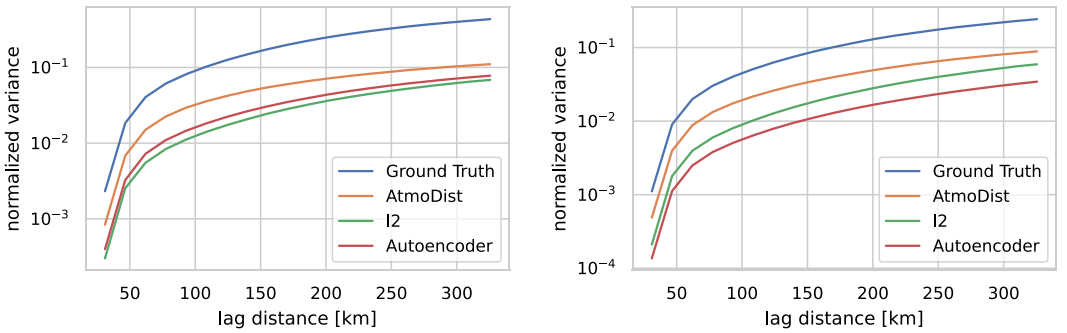

Figure 12. *Semivariograms for the reconstruction of partially occluded fields for divergence (left) and vorticity (right). The autoencoder performs better on divergence than the ℓ_2-loss while the roles are interchanged for vorticity. We hypothesize that this is due to the larger high-frequency content of divergence.*

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

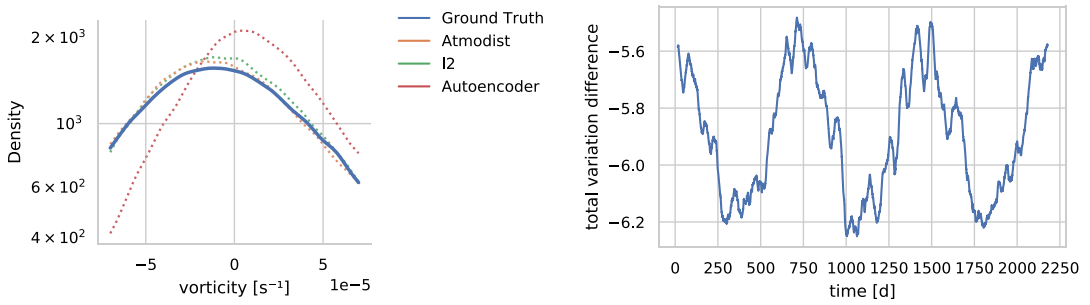

Figure 13. *Left: Kernel density estimate of vorticity distribution at Milan (Italy). The ℓ_2-based GAN achieves a Wasserstein distance of $5.3 \cdot 10^{-6}$ while our approach achieves a Wasserstein distance of $2.0 \cdot 10^{-6}$. The autoencoder-based GAN yields significant worse statistics. Right: Reconstruction error measured as difference of total variation of divergence for the ℓ_2-based super-resolution as a function of time. To highlight the oscillations, the errors have been smoothed by a 30 day moving average. The oscillations are also present in the AtmoDist-based super-resolution, when comparing vorticity, or when the reconstruction error is measured using the ℓ_2 norm.*

the original size. Afterwards, we train three SRGAN models, one for each maximum temporal separation, for 9 epochs using the same hyper-parameters and dataset as in the original downscaling experiment.

Results for the energy spectrum, semivariogram, and downscaling errors are shown in Figure 14. We find that with $\Delta t_{\max} = 69$ hr the downscaling performs slightly better than with $\Delta t_{\max} = 45$ hr with respect to all three metrics. For $\Delta t_{\max} = 93$ hr, the model performs significantly worse than the other two, implying that past a certain threshold, performance begins to degrade. Notably, all three models outperform the ℓ_2-based downscaling model even though the representations networks have been trained with less data as in the main experiment.

5. Conclusion and Future Work

We have presented AtmoDist, a representation learning approach for atmospheric dynamics. It is based on a novel spatiotemporal pretext task designed for atmospheric data that is applicable to a wide range of different fields. We used the representations learned by AtmoDist to introduce a data-driven metric for atmospheric states and showed that it improves the state-of-the-art for downscaling when

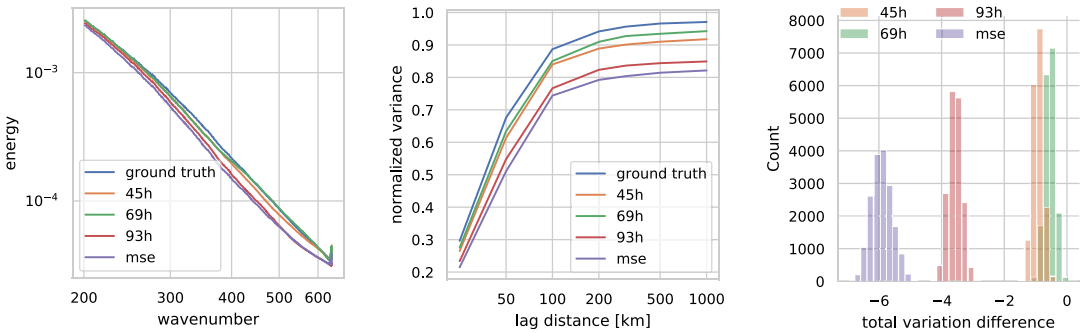

Figure 14. *The energy spectrum (left), semivariogram (center), and distribution of total variation difference errors (right) for models trained with different maximum $\Delta t_{\max}$ for our ablation study. The semivariogram and error distributions are calculated on divergence, but qualitatively similar results are obtained for vorticity.*

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

used as a loss function there. For the reconstruction of missing field data, it also led to more detailed and more realistic results. Surprisingly, AtmoDist improved the performance even for local statistics, although locality played no role in the pretext task. These results validate the quality of our learned representations.

5.1. Possible extensions of AtmoDist

We believe that different extensions of AtmoDist should be explored in the future. One possible direction is the use of a contrastive loss instead of our current pretext task. For this, samples within a certain temporal distance from each other can be used as positive pairs and samples above that threshold as negative ones, akin to word2vec (Mikolov et al., 2013). However, we believe that predicting the exact time lag between two atmospheric states provides a much more challenging task and hence a better training signal than solely predicting if two states are within a certain distance of each other. Exploring a triplet loss (Hoffer and Ailon, 2015) is another interesting direction.

We also want to explore other downstream tasks, for example, the classification and prediction of hurricanes (Prabhat et al., 2021) or extreme events (Racah et al., 2017). Interesting would also be to explore transfer learning for AtmoDist, for example, to train on historical data and then adapt to a regime with significant CO_2 forcing, similar to Barnes et al. (2018). This could be explored with simulation data, which can be used to train AtmoDist without modifications.

We employed only divergence and vorticity and a single vertical layer in AtmoDist. In the future, we want to validate our approach using additional variables, for example, those appearing in the primitive equations, and with more vertical layers. It is also likely that better representations can be obtained when not only a single time step but a temporal window of nearby states is provided to the network.

5.2. Outlook

We consider AtmoDist as a first proof-of-concept for the utility of representation learning for analyzing, understanding, and improving applications in the context of weather and climate dynamics.

Representation learning in computer vision relies heavily on data augmentation (e.g., Chen et al., 2020; Caron et al., 2021). While this is a well-understood subject for natural images, the same does not hold true for atmospheric and more general climate dynamics data. Compared to computer vision, many more physical constraints have to be considered. We hence believe that the design and validation of novel data augmentations is an important direction for future work.

Another currently unexplored research direction is representation learning using (unlabeled) simulation data. For example, one could perform pre-training on the very large amounts of simulation data that are available from Climate Model Intercomparison Project (CMIP) runs (Eyring et al., 2016) and use fine-tuning (Devlin et al., 2019), transfer learning, or domain adaptation to derive a network that is well suited for observational data. Another interesting direction is to compare representations obtained for reanalysis and simulation data, which has the potential to provide insights into subtle biases that persist in simulations.

Our current work focused on improving downstream applications using representation learning. However, we believe that it also has the potential to provide new insights into the physical processes in the atmosphere, analogous to how tools such as proper orthogonal decompositions helped to analyze the physics in the past. In our opinion, in particular attention-based network architectures, such as transformers (Vaswani et al., 2017), provide a promising approach to this.

Acknowledgments. We gratefully acknowledge discussions with the participants of the workshop *Machine Learning and the Physics of Climate* at the Kavli Institute of Theoretical Physics in Santa Barbara that helped to shaped our overall understanding of the potential of representation learning for weather and climate dynamics. Special thanks to Yi Deng for many helpful discussions and explanations.

Author Contributions. Conceptualization: S.H. and C.L. Data curation: S.H. Data visualization: S.H. Methodology: S.H. and C.L. Writing original draft: S.H. and C.L. All authors approved the final submitted draft.

Competing Interests. The authors declare no competing interests exist.

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

Data Availability Statement. Our code is made available at https://github.com/sehoffmann/AtmoDist. Instructions on how to download ERA5 can be found at https://confluence.ecmwf.int/display/CKB/How+to+download+ERA5.

Ethics Statement. The research meets all ethical guidelines, including adherence to the legal requirements of the study country.

Funding Statement. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—Project-ID 422037413–TRR 287.

References

Achille A. and Soatto S (2018) Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, *19*(1):1947–1980.

Aitken A, **Ledig C**, **Theis L**, **Caballero J**, **Wang Z and Shi W** (2017) Checkerboard artifact free sub-pixel convolution:A note on sub-pixel convolution, resize convolution and convolution resize. arXiv preprint arXiv:1707.02937

Arnold VI (1966) Sur la géométrie différentielle des groupes de lie de dimension infinie et ses applications à l'hydrodynamique des fluides parfaits. *Annales de l'institut Fourier 16*, 319–361. Available at http://www.numdam.org/item/AIF_1966__16_1_319_0/.

Ba LJ, **Kiros JR and Hinton GE** (2016) Layer normalization. arXiv preprint arXiv:1607.06450.

Balaji V, **Couvreux F**, **Deshayes J**, **Gautrais J**, **Hourdin F and Rio C** (2022) Are general circulation models obsolete? *Proceedings of the National Academy of Sciences*, *119*(47):e2202075119.

Baldwin MP, **Gray LJ**, **Dunkerton TJ**, **Hamilton K**, **Haynes PH**, **Randel WJ**, **Holton JR**, **Alexander MJ**, **Hirota I**, **Horinouchi T**, **Jones DBA**, **Kinnersley JS**, **Marquardt C**, **Sato K and Takahashi M** (2001) The quasi-biennial oscillation. *Reviews of Geophysics 39*(2), 179–229. https://doi.org/10.1029/1999RG000073. Available at https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/1999RG000073.

Bao H, **Dong L**, **Piao S and Wei F** (2022) BEit: BERT pre-training of image transformers. In *International Conference on Learning Representations*. Available at https://openreview.net/forum?id=p-BhZSz59o4.

Barnes E, **Anderson C and Ebert-Uphoff I** (2018) An ai approach to determining time of emergence of climate change. In *Proceedings of the Eighth International Workshop on Climate Informatics*.

Bengio Y, **Courville A and Vincent P** (2013) Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence 350*(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50.

Bengio Y, **Louradour J**, **Collobert R and Weston J** (2009) Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning, ICML '09*. New York, NY: Association for Computing Machinery, pp. 41–48. https://doi.org/10.1145/1553374.1553380.

Bi K, **Xie L**, **Zhang H**, **Chen X**, **Gu X and Tian Q** (2022) Pangu-weather: A 3d high-resolution model for fast and accurate global weather forecast. arXiv preprint arXiv:2211.02556.

Blanchard A, **Parashar N**, **Dodov B**, **Lessig C and Sapsis T** (2022) A multi-scale deep learning framework for projecting weather extremes. In *NEURIPS 2021 Workshop on Climate Change (Spotlight Talk)*. Available at https://arxiv.org/abs/2210.12137.

Brown T, **Mann B**, **Ryder N**, **Subbiah M**, **Kaplan JD**, **Dhariwal P**, **Neelakantan A**, **Shyam P**, **Sastry G**, **Askell A**, **Agarwal S**, **Herbert-Voss A**, **Krueger G**, **Henighan T**, **Child R**, **Ramesh A**, **Ziegler D**, **Wu J**, **Winter C**, **Hesse C**, **Chen M**, **Sigler E**, **Litwin M**, **Gray S**, **Chess B**, **Clark J**, **Berner C**, **McCandlish S**, **Radford A**, **Sutskever I and Amodei D** (2020) Language models are few-shot learners. In Larochelle H, Ranzato M, Hadsell R, Balcan M and Lin H (eds), *Advances in Neural Information Processing Systems*, Vol. *33*. Red Hook, NY: Curran Associates, Inc., pp. 1877–1901. Available at https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Caron M, **Touvron H**, **Misra I**, **Jegou H**, **Mairal J**, **Bojanowski P and Joulin A** (2021) Emerging properties in self-supervised vision transformers. In Proceedings of the IEEE/CVF international conference on computer vision, pages 9650–9660.

Chen T, **Kornblith S**, **Norouzi M and Hinton GE** (2020) A simple framework for contrastive learning of visual representations. In Iii HD and Singh A (eds), *Proceedings of the 37th International Conference on Machine Learning*, Vol. *119*. PMLR, pp. 1597–1607. Available at https://proceedings.mlr.press/v119/chen20j.html.

Chicco D (2021) *Siamese Neural Networks: An Overview*. New York, NY: Springer, pp. 73–94. https://doi.org/10.1007/978-1-0716-0826-5_3.

Devlin J, **Chang M-W**, **Lee K and Toutanova K** (2019) Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv preprint arXiv:1810.04805.

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** (2020) An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929

Dueben PD and Bauer P (2018) Challenges and design choices for global weather and climate models based on machine learning. *Geoscientific Model Development 11*(10), 3999–4009. https://doi.org/10.5194/gmd-11-3999-2018; Available at https://gmd.copernicus.org/articles/11/3999/2018/.

Ebin DG and Marsden JE (1970) Groups of diffeomorphisms and the motion of an incompressible fluid. *The Annals of Mathematics 92*(1), 102–163. Available at http://www.jstor.org/stable/1970699.

Eyring V, **Bony S**, **Meehl GA**, **Senior CA**, **Stevens B**, **Stouffer RJ and Taylor KE** (2016) Overview of the coupled model intercomparison project phase 6 (cmip6) experimental design and organization. *Geoscientific Model Development 9*(5), 1937–1958. https://doi.org/10.5194/gmd-9-1937-2016; Available at https://gmd.copernicus.org/articles/9/1937/2016/.

Gatys LA, **Ecker AS and Bethge M** (2016) Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gidaris S, **Singh P and Komodakis N** (2018) Unsupervised representation learning by predicting image rotations. In *ICLR 2018, Vancouver, Canada*. Available at https://hal-enpc.archives-ouvertes.fr/hal-01864755.

Goodfellow IJ, **Pouget-Abadie J**, **Mirza M**, **Xu B**, **Warde-Farley D**, **Ozair S**, **Courville A and Bengio Y** (2014) Generative adversarial networks. In *Proceedings of the International Conference on Neural Information Processing Systems*, pp. 2672–2680.

Groenke B, **Madaus L and Monteleoni C** (2020) Climalign: Unsupervised statistical downscaling of climate variables via normalizing flows. In *Proceedings of the 10th International Conference on Climate Informatics, CI2020*. New York, NY: Association for Computing Machinery, pp. 60–66. https://doi.org/10.1145/3429309.3429318.

Hannachi A and Iqbal W (2019) Bimodality of hemispheric winter atmospheric variability via average flow tendencies and kernel EOFs. *Tellus A: Dynamic Meteorology and Oceanography 71*(1), 1633847. https://doi.org/10.1080/16000870.2019.1633847.

He K, **Chen X**, **Xie S**, **Li Y**, **Dollár P and Girshick R** (2022) Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009.

He K, **Fan H**, **Wu Y**, **Xie S and Girshick R** (2020) Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9729–9738.

He K, **Zhang X**, **Ren S and Sun J** (2015) Deep residual learning for image recognition. arXiv preprint arXiv:1512.03385.

Hersbach H, **Bell B**, **Berrisford P**, **Hirahara S**, **Horányi A**, **Muñoz-Sabater J**, **Nicolas J**, **Peubey C**, **Radu R**, **Schepers D**, **Simmons A**, **Soci C**, **Abdalla S**, **Abellan X**, **Balsamo G**, **Bechtold P**, **Biavati G**, **Bidlot J**, **Bonavita M**, **De Chiara G**, **Dahlgren P**, **Dee D**, **Diamantakis M**, **Dragani R**, **Flemming J**, **Forbes R**, **Fuentes M**, **Geer A**, **Haimberger L**, **Healy S**, **Hogan RJ**, **Hólm E**, **Janisková M**, **Keeley S**, **Laloyaux P**, **Lopez P**, **Lupu C**, **Radnoti G**, **de Rosnay P**, **Rozum I**, **Vamborg F**, **Villaume S and Thépaut J-N** (2020) The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society 146*(730), 1999–2049. https://doi.org/10.1002/qj.3803; Available at https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/qj.3803.

Hoffer E and Ailon N (2015) Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*. Cham: Springer, pp. 84–92.

Ioffe S (2017) Batch renormalization: towards reducing minibatch dependence in batch-normalized models. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 1942–1950.

Jean N, **Wang S**, **Samar A**, **Azzari G**, **Lobell D and Ermon S** (2019) Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33, pp. 3967–3974.

Jiang CM, **Esmaeilzadeh S**, **Azizzadenesheli K**, **Kashinath K**, **Mustafa M**, **Tchelepi HA**, **Marcus P**, **Prabhat M and Anandkumar A** (2020) *MeshfreeFlowNet: A Physics-Constrained Deep Continuous Space-Time Super-Resolution Framework*. IEEE Press.

Karras T, **Laine S and Aila T** (2019) A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4401–4410. Available at https://arxiv.org/abs/1812.04948.

Karras T, **Laine S**, **Aittala M**, **Hellsten J**, **Lehtinen J and Aila T** (2020) Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8110–8119. Available at https://arxiv.org/abs/1912.04958.

Klemmer K and Neill DB (2021) Auxiliary-task learning for geographic data with autoregressive embeddings. In *Proceedings of the 29th International Conference on Advances in Geographic Information Systems, SIGSPATIAL '21*. New York, NY: Association for Computing Machinery, pp. 141–144. https://doi.org/10.1145/3474717.3483922.

Klemmer K, **Xu T**, **Acciaio B and Neill DB** (2022) Spate-gan: Improved generative modeling of dynamic spatio-temporal patterns with an autoregressive embedding loss. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36, pp. 4523–4531. https://doi.org/10.1609/aaai.v36i4.20375; Available at https://ojs.aaai.org/index.php/AAAI/article/view/20375.

Koh T-Y and Wan F (2015) Theory of the norm-induced metric in atmospheric dynamics. *Atmospheric Chemistry and Physics 150* (5), 2571–2594. https://doi.org/10.5194/acp-15-2571-2015; Available at https://acp.copernicus.org/articles/15/2571/2015/.

Kurinchi-Vendhan R, **Lütjens B**, **Gupta R**, **Werner L and Newman D** (2021) WiSoSuper: Benchmarking Super-Resolution Methods on Wind and Solar Data. arXiv preprint arXiv:2109.08770.

Lam R, **Sanchez-Gonzalez A**, **Willson M**, **Wirnsberger P**, **Fortunato M**, **Pritzel A**, **Ravuri S**, **Ewalds T**, **Alet F**, **Eaton-Rosen Z**, **Hu W**, **Merose A**, **Hoyer S**, **Holland G**, **Stott J**, **Vinyals O**, **Mohamed S and Battaglia P** (2022) GraphCast: Learning skillful medium-range global weather forecasting. arXiv preprint arXiv:2212.12794.

LeCun Y, **Bottou L**, **Bengio Y and Haffner P** (1998) Gradient-based learning applied to document recognition. *Proceedings of the IEEE 860*(11), 2278–2324.

Ledig C, **Theis L**, **Huszar F**, **Caballero J**, **Cunningham A**, **Acosta A**, **Aitken A**, **Tejani A**, **Totz J**, **Wang Z and Shi W** (2017) Photo-realistic single image super-resolution using a generative adversarial network. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Le-Khac PH, **Healy G and Smeaton AF** (2020) Contrastive representation learning: A framework and review. *IEEE Access 80*, 193907–193934. https://doi.org/10.1109/ACCESS.2020.3031549.

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

Lim B, **Son S**, **Kim H**, **Nah S and Lee KM** (2017) Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 136–144.

Lima CH, **Lall U**, **Jebara T and Barnston AG** (2009) Statistical prediction of enso from subsurface sea temperature using a nonlinear dimensionality reduction. *Journal of Climate 22*(17), 4501–4519.

Lorenz EN (1969) The predictability of a flow which possesses many scales of motion. *Tellus 21*(3), 289–307. https://doi.org/10.1111/j.2153-3490.1969.tb00444.x; Available at https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2153-3490.1969.tb00444.x.

Matheron G (1963) Principles of geostatistics. *Economic Geology 58*(8), 1246–1266.

Meraner A, **Ebel P**, **Zhu XX and Schmitt M** (2020) Cloud removal in sentinel-2 imagery using a deep residual neural network and Sar-optical data fusion. *ISPRS Journal of Photogrammetry and Remote Sensing 1660*, 333–346. https://doi.org/10.1016/j.isprsjprs.2020.05.013; Available at https://www.sciencedirect.com/science/article/pii/S0924271620301398.

Mercer AE and Richman MB (2012) Assessing atmospheric variability using kernel principal component analysis. *Procedia Computer Science 12*, 288–293. https://doi.org/10.1016/j.procs.2012.09.071; Available at https://www.sciencedirect.com/science/article/pii/S187705091200662X.

Mikolov T, **Chen K**, **Corrado G and Dean J** (2013) Efficient estimation of word representations in vector space. In Bengio Y and LeCun Y (eds), *1st International Conference on Learning Representations, ICLR, Scottsdale, Arizona, USA, May 2–4, 2013, Workshop Track Proceedings*. Available at http://arxiv.org/abs/1301.3781.

Misra I, **Zitnick CL and Hebert M** (2016) Shuffle and learn: Unsupervised learning using temporal order verification. In Leibe B, Matas J, Sebe N and Welling M (eds), *Computer Vision - ECCV - 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I, Volume 9905 of Lecture Notes in Computer Science*. Springer, pp 527–544. https://doi.org/10.1007/978-3-319-46448-0_32.

Noroozi M and Favaro P (2016) Unsupervised learning of visual representations by solving jigsaw puzzles. In Leibe B, Matas J, Sebe N and Welling M (eds), *Computer Vision – ECCV 2016*. Cham: Springer International Publishing, pp 69–84.

Palmer TN, **Gelaro R**, **Barkmeijer J and Buizza R** (1998) Singular vectors, metrics, and adaptive observations. *Journal of the Atmospheric Sciences 55*(4):633–653. https://doi.org/10.1175/1520-0469(1998)055<0633:SVMAAO>2.0.CO;2; Available at https://journals.ametsoc.org/view/journals/atsc/55/4/1520-0469_1998_055_0633_svmaao_2.0.co_2.xml.

Pathak D, **Krahenbuhl P**, **Donahue J**, **Darrell T and Efros AA** (2016) Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Pathak J, **Subramanian S**, **Harrington P**, **Raja S**, **Chattopadhyay A**, **Mardani M**, **Kurth T**, **Hall D**, **Li Z**, **Azizzadenesheli K**, **Hassanzadeh P**, **Kashinath K and Anandkumar A** (2022) FourCastNet: A Global Data-driven High-resolution Weather Model using Adaptive Fourier Neural Operators. arXiv preprint arXiv:2202.11214.

Prabhat K, **Kashinath K**, **Mudigonda M**, **Kim S**, **Kapp-Schwoerer L**, **Graubner A**, **Karaismailoglu E**, **von Kleist L**, **Kurth T**, **Greiner A**, **Mahesh A**, **Yang K**, **Lewis C**, **Chen J**, **Lou A**, **Chandran S**, **Toms B**, **Chapman W**, **Dagon K**, **Shields CA**, **O'Brien T**, **Wehner M and Collins W** (2021) Climatenet: An expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. *Geoscientific Model Development 14*(1):107–124. https://doi.org/10.5194/gmd-14-107-2021; Available at https://gmd.copernicus.org/articles/14/107/2021/.

Racah E, **Beckham C**, **Maharaj T**, **Kahou S**, **Prabhat M and Pal C** (2017) Extremeweather: A large-scale climate dataset for semi-supervised detection, localization, and understanding of extreme weather events. In Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S and Garnett R (eds), *Advances in Neural Information Processing Systems 30*. Red Hook, NY: Curran Associates, Inc., pp. 3405–3416; Available at http://papers.nips.cc/paper/6932-extremeweather-a-large-scale-climate-dataset-for-semi-supervised-detection-localization-and-understanding-of-extreme-weather-events.pdf.

Radford A, **Metz L and Chintala S** (2015) Unsupervised representation learning with deep convolutional generative adversarial networks. arXiv preprint arXiv:1511.06434.

Radford A, **Narasimhan K**, **Salimans T and Sutskever I** (2018) Improving language understanding by generative pre-training.

Ranftl R, **Bochkovskiy A and Koltun V** (2021) Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12179–12188.

Rasp S, **Dueben PD**, **Scher S**, **Weyn JA**, **Mouatadid S and Thuerey N** (2020) Weatherbench: A benchmark data set for data-driven weather forecasting. *Journal of Advances in Modeling Earth Systems 12*(11), e2020MS002203.

Reichstein M, **Camps-Valls G**, **Stevens B**, **Jung M**, **Denzler J**, **Carvalhais N and Prabhat** (2019) Deep learning and process understanding for data-driven earth system science. *Nature 566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1; Available at http://www.nature.com/articles/s41586-019-0912-1.

Requena-Mesa C, **Reichstein M**, **Mahecha M**, **Kraft B and Denzler J** (2019) Predicting landscapes from environmental conditions using generative networks. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 11824 LNCS, pp. 203–217. https://doi.org/10.1007/978-3-030-33676-9_14; Available at https://www.scopus.com/inward/record.uri?eid=2-s2.0-85076136395.

Ronneberger O, **Fischer P and Brox T** (2015) U-net: Convolutional networks for biomedical image segmentation. In Navab N, Hornegger J, Wells WM and Frangi AF (eds), *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*. Cham: Springer International Publishing, pp. 234–241.

Russakovsky O, **Deng J**, **Su H**, **Krause J**, **Satheesh S**, **Ma S**, **Huang Z**, **Karpathy A**, **Khosla A**, **Bernstein M**, **Berg AC and Fei-Fei L** (2015) ImageNet Large Scale Visual Recognition Challenge. *International journal of computer vision*, *115*, 211–252.

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

Schultz M, **Betancourt C**, **Gong B**, **Kleinert F**, **Langguth M**, **Leufen LH**, **Mozaffari A and Stadtler S** (2021) Can deep learning beat numerical weather prediction? *Philosophical Transactions of the Royal Society of London/A 379*(2194), 20200097. https://doi.org/10.1098/rsta.2020.0097; Available at https://juser.fz-juelich.de/record/890552.

Simonyan K and Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In *ICLR 2015*.

Stengel K, **Glaws A**, **Hettinger D and King RN** (2020) Adversarial super-resolution of climatological wind and solar data. *Proceedings of the National Academy of Sciences 117*(29), 16805–16815. https://doi.org/10.1073/pnas.1918964117; Available at https://www.pnas.org/content/117/29/16805.

Talagrand O (1981) A study of the dynamics of four-dimensional data assimilation. *Tellus 33*(1), 43–60. https://doi.org/10.3402/tellusa.v33i1.10693.

Tolstikhin I, **Houlsby N**, **Kolesnikov A**, **Beyer L**, **Zhai X**, **Unterthiner T**, **Yung J**, **Steiner A**, **Keysers D**, **Uszkoreit J**, **Lucic M and Dosovitskiy A** (2021) Mlp-mixer: An all-mlp architecture for vision. *Advances in neural information processing systems*, *34*, 24261–24272.

Toms BA, **Barnes EA and Ebert-Uphoff I** (2020) Physically interpretable neural networks for the geosciences: Applications to earth system variability. *Journal of Advances in Modeling Earth Systems 12*(9), e2019MS002002. https://doi.org/10.1029/2019MS002002; Available at https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS002002.

Ulyanov D, **Vedaldi A and Lempitsky V** (2017) Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6924–6932.

Vaswani A, **Shazeer N**, **Parmar N**, **Uszkoreit J**, **Jones L**, **Gomez AN**, **Kaiser L and Polosukhin I** (2017) Attention is all you need. In Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S and Garnett R (eds), *Advances in Neural Information Processing Systems*, Vol. *30*. Red Hook, NY: Curran Associates, Inc.. Available at https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Vincent P, **Larochelle H**, **Lajoie I**, **Bengio Y and Manzagol P-A** (2010) Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of Machine Learning Research 11*(110), 3371–3408. Available at http://jmlr.org/papers/v11/vincent10a.html.

Wang Z, **Bovik AC**, **Sheikh HR and Simoncelli EP** (2004) Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing 13*(4), 600–612.

Wang X, **Yu K**, **Wu S**, **Gu J**, **Liu Y**, **Dong C**, **Qiao Y and Change Loy C** (2018) Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*.

Watson-Parris D, **Rao Y**, **Olivié D**, **Seland Ø**, **Nowack P**, **Camps-Valls G**, **Stier P**, **Bouabid S**, **Dewey M**, **Fons E**, **Gonzalez J**, **Harder P**, **Jeggle K**, **Lenhardt J**, **Manshausen P**, **Novitasari M**, **Ricard L and Roesch C** (2022) Climatebench v1. 0: A benchmark for data-driven climate projections. *Journal of Advances in Modeling Earth Systems 14*(10), e2021MS002954.

Wieczorek MA and Meschede M (2018) Shtools: Tools for working with spherical harmonics. *Geochemistry, Geophysics, Geosystems 19*(8), 2574–2592. https://doi.org/10.1029/2018GC007529; Available at https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2018GC007529.

Yang C, **Lu X**, **Lin Z**, **Shechtman E**, **Wang O and Li H** (2017) High-resolution image inpainting using multi-scale neural patch synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21–26, 2017*. IEEE Computer Society, pp. 4076–4084. https://doi.org/10.1109/CVPR.2017.434.

Zeiler MD and Fergus R (2014) Visualizing and understanding convolutional networks. In Fleet D, Pajdla T, Schiele B and Tuytelaars T (eds), *Computer Vision – ECCV 2014*. Cham: Springer International Publishing, pp. 818–833.

Zhai X, **Kolesnikov A**, **Houlsby N and Beyer L** (2021) Scaling Vision Transformers. arXiv preprint arXiv:2106.04560.

Zhang R, **Isola P and Efros AA** (2017) Split-brain autoencoders: Unsupervised learning by cross-channel prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang R, **Isola P**, **Efros AA**, **Shechtman E and Wang O** (2018) The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 586–595).

Zhu D, **Cheng X**, **Zhang F**, **Yao X**, **Gao Y and Liu Y** (2020) Spatial interpolation using conditional generative adversarial neural networks. *International Journal of Geographical Information Science 34*(4), 735–758.

A. Appendix.

A.1. Preprocessing. Divergence and vorticity are transformed in a preprocessing step by $y = f(g(h(x)))$ where

$$y = f(w) = \frac{w-\mu_2}{\sigma_2} \quad w = g(z) = \operatorname{sign}(z)\log(1+\alpha|z|) \quad z = h(x) = \frac{x-\mu_1}{\sigma_1} \tag{A.1}$$

and which is applied element-wise and independently for vorticity and divergence. Here μ_1 and σ_1 denote the mean and standard deviation of the corresponding fields, respectively, while μ_2 and σ_2 denote the mean and standard deviation of the log-transformed field w. All moments are calculated across the training dataset and are shown in Table 3. The parameter α controls the strength by which the dynamic range at the tails of the distribution is compressed. We found that $\alpha = 0.2$ is sufficient to stabilize training while it avoids an aggressive compression of the original data. Notice that the log function behaves approximately linear around 1, thus leaving small values almost unaffected.

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

Table 3. Mean and standard deviations calculated on the training dataset (1979–1998) on model level 120 for divergence and relative vorticity.

Variable	μ_1	σ_1	μ_2	σ_2
Divergence	1.9464×10^{-8}	2.8569×20^{-5}	8.821×10^{-4}	1.5795×10^{-1}
(Rel.) Vorticity	2.0548×10^{-7}	5.0819×10^{-5}	3.2483×10^{-4}	1.6044×10^{-1}

A.2. AtmoDist Training. The AtmoDist network is trained using standard stochastic gradient descent with momentum $\beta = 0.9$ and an initial learning rate of $\eta = 10^{-1}$. If training encounters a plateau, the learning rate is reduced by an order of magnitude to a minimum of $\eta_{\min} = 10^{-5}$. Additionally, gradient clipping is employed, ensuring that the l_2-norm of the gradient does not exceed $G_{\max} = 5.0$. Finally, to counteract overfitting, weight decay of 10^{-4} is used.

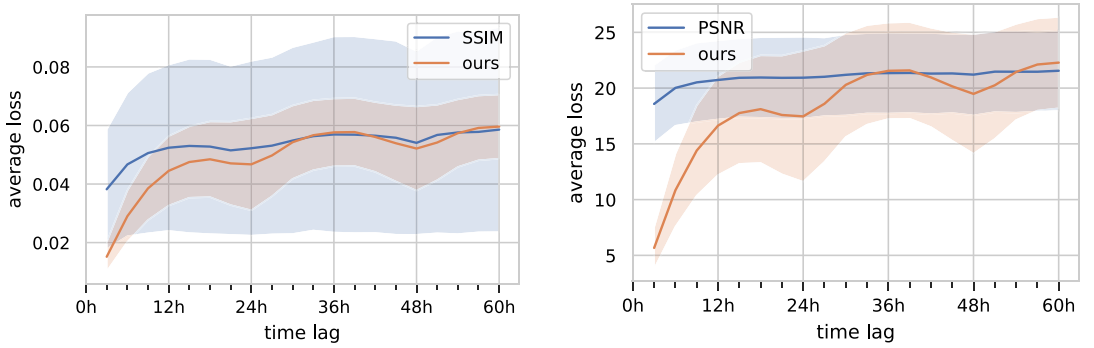

Figure 15. *Mean SSIM and PSNR as a function of the temporal separation Δt. Since in both cases higher quantities indicate more similarity between samples, we apply the following transformations to make the plots comparable to Figure 4: SSIM: $y = 1 - (1 + SSIM(X_{t_1}, X_{t_2}))/2$; PSNR: $y = 50dB - PSNR(X_{t_1}, X_{t_2})$.*

Despite the network converging on lower resolutions in preliminary experiments, once we trained on 160x160 patches at native resolution (1280x2560) the network failed to converge. We hypothesize that the issue is the difficulty of the pretext task combined with an initial lack of discerning features. We thus employ a pre-training scheme inspired by curriculum learning (Bengio et al., 2009). More specifically, we initially train the network only on about 10% of the data so that it can first focus on solving the task there. After 20 epochs, we then reset the learning rate to $\eta = 10^{-1}$ and start training on the whole dataset.

A.3. Scaling the Loss Function. To ensure that downscaling with ℓ_2-loss and the AtmoDist metric exhibit the same training dynamics, we normalize our loss function. This is particularly important with respect to the α_{adv} parameter which controls the trade-off between content-loss and adversarial-loss in SRGAN (Ledig et al., 2017). The same procedure is also applied to the loss function derived from the autoencoder.

We hypothesize that due to the chaotic dynamics of the atmosphere, any loss function should on average converge to a specific level after a certain time period (ignoring daily and annual oscillations). Thus, we normalize our content-loss by ensuring that the equilibrium levels are roughly the same in terms of least squares by solving the following optimization problem for the scaling factor α

$$\widehat{\alpha} = \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} \sum_{t=\frac{N}{2}}^{N} (\alpha c_t - m_t)^2 \tag{A.2}$$

where c_t denote the average AtmoDist distance of samples that are Δt apart and m_t their average ℓ_2 distance. It is easy to verify that the above optimization problem has the unique solution

$$\widehat{\alpha} = \frac{\sum_{t=\frac{N}{2}}^{N} c_t m_t}{\sum_{t=\frac{N}{2}}^{N} c_t^2}. \tag{A.3}$$

A.4. Autoencoder Architecture. The autoencoder takes as input a divergence and vorticity field of size 160×160. It consists of an encoder part that compresses the input field to a suitable representation, and a decoder that takes the representation and reconstructs the original field from it.

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

The encoder is identical to the representation network used for the AtmoDist task. The decoder is a mirrored version of the encoder where downscaling convolutions were replaced by upscaling ones. Upscaling is done by bilinear interpolation followed by a standard residual block.

In the middle, that is in-between encoder and decoder, no feed-forward layer is used. It would have contained the majority of parameters of the overall network and thus potentially also of its capacity. Instead, we use an approach inspired by Tolstikhin et al. (2021) to ensure that information can propagate between each spatial position.

First, the $H \times \mathrm{W} \times C$ feature map of the last encoder layer, where H, W, C denote height, width, and number of channels, respectively, is interpreted as C vectors of dimensionality $H \cdot W$. Each vector is then transformed by a feed-forward layer, mixing information spatially for each channel. Afterwards, the same procedure is repeated for the $H \cdot W$ vectors of dimensionality C to mix information channel-wise as well. This approach allows us to propagate information globally without bloating the size of the network in a significant way.

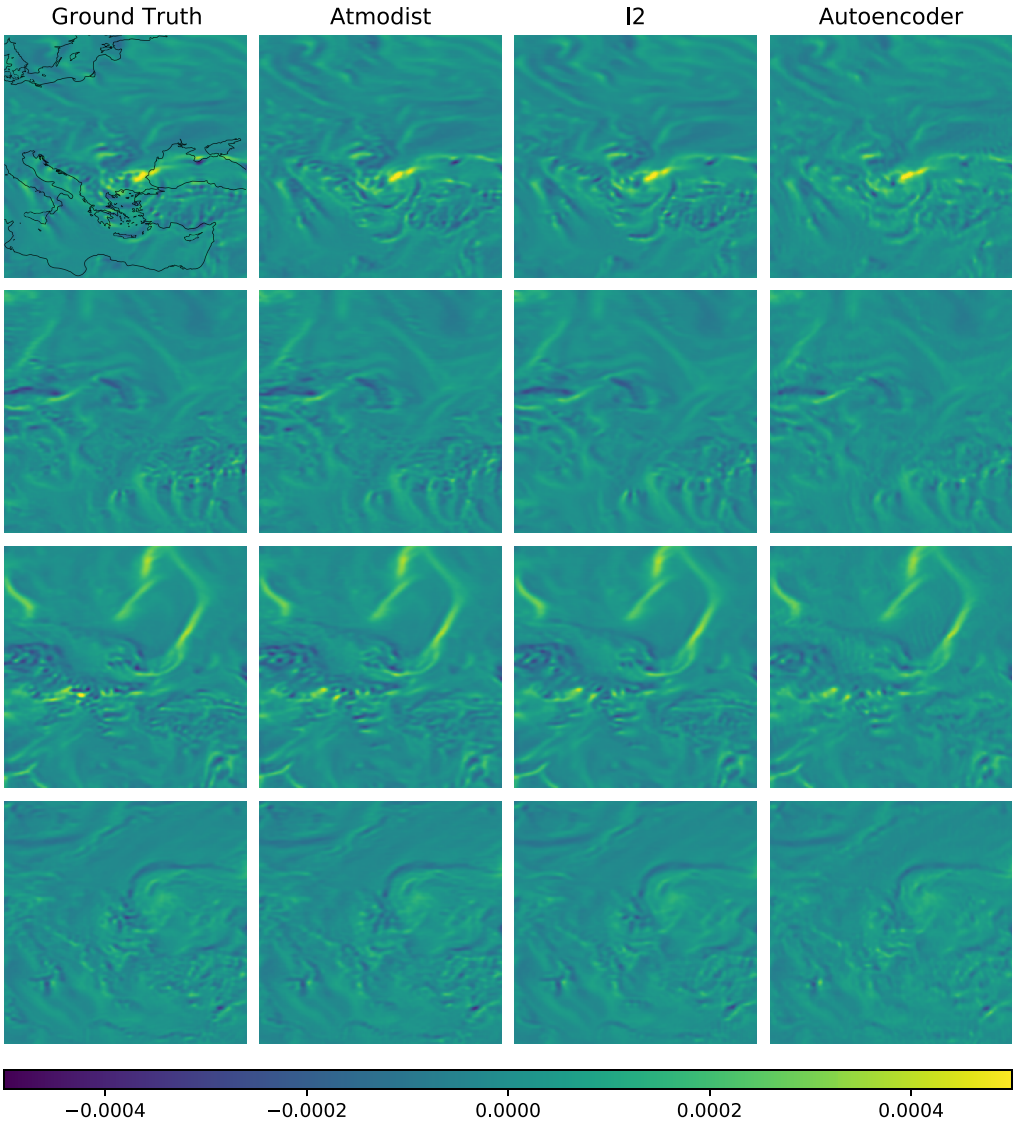

Figure 16. *Uncurated set of downscaled vorticity fields over the Mediterranean Sea and Eastern Europe at different time steps. Coastlines are shown in the first ground truth field and then omitted for better comparability.*

https://doi.org/10.1017/eds.2023.1 Published online by Cambridge University Press

The autoencoder is trained in the same way as the AtmoDist representation network, compare Appendix A.2, except that no pre-training is used.

A.5. Inpainting Training. Our network used for inpainting follows the same architecture as the autoencoder described above.

When training with either the AtmoDist- or autoencoder-based loss function, we initialize the network with a pre-trained version in the same way as we did for the super-resolution already. Furthermore, a small $l2$-loss term is added as a regularization. Specifically, the total loss is given by

$$L(X_1, X_2) = (1-\gamma) L_{\text{content}}(X_1, X_2) + \gamma \|X_1 - X_2\|_2, \tag{A.4}$$

where $\gamma = 0.1$ and L_{content} is either the already scaled AtmoDist or autoencoder loss. This is done to prevent local minima and artifacts during training.

A.6. Semivariogram Calculation. The semivariogram, given by

$$\gamma(r) = \int (f(x+r) - f(x))^2 \mathrm{d}x, \tag{A.5}$$

can be calculated in different ways. We approximate the integral that defines it using Monte-Carlo sampling. In particular, for each time-step and each lag-distance r, 300 random locations and 300 random directions are sampled and the field is evaluated at these points. This procedure is done for the complete evaluation period and in the end the semivariogram is obtained by averaging.

Cite this article: Hoffmann S. and Lessig C. (2023). AtmoDist: Self-supervised representation learning for atmospheric dynamics. *Environmental Data Science*, 2: e6. doi:10.1017/eds.2023.1