Multivariate Emulation of Kilometer-Scale Numerical Weather Predictions with Generative Adversarial Networks: A Proof of Concept

Clément Brochet, Laure Raynaud, Nicolas Thome, Matthieu Plu, Clément Rambour

Multivariate emulation of kilometer-scale numerical weather predictions with generative adversarial networks: a proof-of-concept

Clément Brochet[a,b] , Laure Raynaud [b] , Nicolas Thome [c] , Matthieu Plu [b] , Clément Rambour [c]

[a] *Ecole des Ponts Paris-Tech, France,* [b] *CNRM, Université de Toulouse, Météo-France, CNRS, Toulouse, France,* [c] *CNAM, France*

Corresponding author: Clément Brochet, clement.brochet@meteo.fr

Early Online Release: This preliminary version has been accepted for publication in *Artificial Intelligence for the Earth Systems*, may be fully cited, and has been assigned DOI 10.1175/AIES-D-23-0006.1. The final typeset copyedited article will replace the EOR at the above DOI when it is published.

ABSTRACT: Emulating numerical weather prediction (NWP) model outputs is important to compute large datasets of weather fields in an efficient way. The purpose of the present paper is to investigate the ability of generative adversarial networks (GAN) to emulate distributions of multivariate outputs (10-meter wind and 2-meter temperature) of a kilometer-scale NWP model. For that purpose, a residual GAN architecture, regularized with spectral normalization, is trained against a kilometer-scale dataset from the AROME ensemble prediction system (AROME-EPS). A wide range of metrics is used for quality assessment, including pixel-wise and multi-scale earth-mover distances, spectral analysis, and correlation length scales. The use of wavelet-based scattering coefficients as meaningful metrics is also presented. The GAN generates samples with good distribution recovery and good skill in average spectrum reconstruction. Important local weather patterns are reproduced with a high level of detail, while the joint generation of multivariate samples matches the underlying AROME-EPS distribution. The different metrics introduced describe the GAN's behavior in a complementary manner, highlighting the need to go beyond spectral analysis in generation quality assessment. An ablation study then shows that removing variables from the generation process is globally beneficial, pointing at the GAN limitations to leverage cross-variable correlations. The role of absolute positional bias in the training process is also characterized, explaining both accelerated learning and quality-diversity trade-off in the multivariate emulation. These results open perspectives about the use of GAN to enrich NWP ensemble approaches, provided that the aforementioned positional bias is properly controlled.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

1. Introduction

Having access to large sets of weather forecasts or reforecasts is of plain importance in many applications. For instance, some fundamental and applied studies in weather science rely on large reforecasts of events, e.g., to detect climatological trends on specific patterns such as extratropical depressions (Pantillon et al. 2017) or heavy precipitating events (Ponzano et al. 2020). Such reforecasts, like operational forecasts in many centers, are usually based on ensemble prediction systems (EPSs). However, the high computing and storage costs of these systems can limit their configuration to a few dozen of members for convection-permitting, kilometer-scale ensembles. This is often not enough to accurately sample the future distributions of weather variables. As a consequence, emulation of forecasts at different scales can have potential applications for both climatological studies and operational weather forecasting.

Increasing the ensemble size without resorting to the costly option of running additional EPS members remains an open challenge. Existing solutions mainly rely on different flavours of neighbourhood approaches (Roberts and Lean 2008; Ebert 2008), that are based on the assumption of locally homogeneous weather. An original approach has been proposed by Vincendon et al. (2011) to design perturbed precipitation forecasts by applying location and intensity perturbations to the deterministic forecast. In recent years, generative deep learning has emerged as a novel approach that can produce accurate synthetic data, and it has recently seen a broadening use by the NWP community. In particular, generative adversarial networks (GAN, Goodfellow et al. (2014)) and variational auto-encoders (VAE, Kingma and Welling (2014)) have already been used for several NWP applications (Bihlo 2020; Ravuri et al. 2021; Leinonen et al. 2021; Bhatia et al. 2021; Harris et al. 2022). Of particular interest is the recent work of Besombes et al. (2021), that demonstrates the ability of a GAN to emulate realistic atmospheric states (accounting for several variables at different atmospheric levels) when trained on outputs of a simple climate model at a relatively coarse resolution ($\approx 300km$).

Generative techniques allow sample draws from a simple, latent distribution, which are then mapped into higher-dimensional spaces (e.g the space of NWP model outputs). VAEs have an encoder-latent-decoder structure: their latent space is used both to embed the input samples, and then to generate maximum-likelihood samples from a parameterized distribution. This set-up is prone to creating noisy samples, as exemplified by Dumoulin et al. (2016), and the de-noised

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

samples can be blurry (Kingma and Welling 2019). GANs on the other hand are composed of two competing networks (named the generator and the discriminator). Once trained, the generator of a GAN usually produces highly detailed images (Radford et al. 2015; Karras et al. 2018).

Although the performances of GANs can be appealing, it can be challenging to stably train a GAN model (Goodfellow et al. 2014; Arjovsky et al. 2017), as they are commonly affected by several obstacles. The first one is *mode collapse*, which is the concentration of the samples produced towards a small portion of the training distribution, and in extreme cases, a single sample. This can be the case if the generator begins to reproduce a specific subset of the training set to anomalously sharp numerical precision (Radford et al. 2015). It is a kind of overfitting. A second difficulty is the sudden loss of quality of the generated samples. This can be due to inefficient feature extraction from the discriminator, or from discriminator's overfitting (Brock et al. 2018). This polarity between samples quality and distribution recovery has been termed the *quality-diversity trade-off* (Brock et al. 2018), or more recently the *perception-distortion trade-off* (Blau and Michaeli 2018). Therefore, evaluating a GAN should focus on two main aspects: the intrinsic quality of the samples, and the recovery of the main features of the training dataset ; this requires specific metrics.

Building on the work of Besombes et al. (2021), the objective of the present article is to examine the ability of GANs to emulate multivariate outputs of NWP models at a kilometric resolution, close to the one studied by Ravuri et al. (2021). To the authors' knowledge, this aspect has not been evaluated yet, and the sensitivity of GAN training advocates for a dedicated study. Two important questions will be addressed: are GANs effectively able to emulate multivariate outputs with a proper representation of every spatial scale ? How can one evaluate the diversity and realism of the outputs of a GAN trained on such data ? This is a preliminary step before using GANs to enhance EPSs, although such a task is left for future work.

This study proposes the training of a residual, spectrally normalized Wasserstein-GAN (Miyato et al. 2018), using kilometer-scale model outputs from the AROME Ensemble Prediction System (AROME-EPS). The AROME-EPS dataset involves several fields exhibiting fine-scale variations, such as 10-meter wind speed components and 2-meter temperature. To analyze what effect different variables have on training, several configurations will be examined with distinct sets of variables. Borrowing from weather science and computer vision, a comprehensive set of metrics is considered to assess different aspects of the GAN's outputs. The spatial structure of emulated fields is evaluated

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

with spectral transforms, correlation length scales, and scattering coefficients. These metrics are complemented with distributional distances, using pixel-wise Wasserstein distance (Besombes et al. 2021) and sliced Wasserstein distance (Rabin et al. 2011; Karras et al. 2018). With this set of metrics, a detailed view of the GAN's capabilities and weaknesses is provided, and we assess its sensitivity to the choice of hyperparameters and to the chosen architecture and set-up.

The outline of the paper is as follows. The dataset and choices made for the setup are detailed in Section 2; this includes choice of data, network architecture, implementation of the GAN training algorithm. Section 3 details the whole set of metrics to be used in evaluation, and the evaluation strategy. Section 4 presents the main results obtained for the joint emulation of three AROME variables with a GAN. Section 5 compares the results obtained when varying the number and nature of the AROME variables used as predictors. Section 6 discusses the results. Section 7 provides conclusions and opens some perspectives for future work.

2. Generating AROME forecasts with a GAN: setup choices

a. Dataset and problem formulation

The dataset used is made of forecasts from the 16-member, 1.3 km resolution AROME Ensemble Prediction System (AROME-EPS) (Bouttier et al. 2012; Raynaud and Bouttier 2016), covering about 17 continuous months, from 15 June 2020 up to 12 November 2021. AROME-EPS is the ensemble version of the AROME limited-area, convection-permitting model (Seity et al. 2011), used operationally at Météo-France. The version of AROME-EPS considered uses a 1.3 km grid resolution, and produces outputs on a regular latitude-longitude grid at 0.025° resolution. Initial conditions are built using the AROME 3D-Var analysis (Brousseau et al. 2011) and perturbations from the AROME Ensemble Data Assimilation (Montmerle et al. 2018), while lateral boundary conditions are given by forecasts of the global ARPEGE-EPS model (Descamps et al. 2015). AROME-EPS also uses the stochastically-perturbed parametrisation tendencies (SPPT) scheme (Bouttier et al. 2012) and surface perturbations (Bouttier et al. 2016).

AROME-EPS 16-member forecasts are launched daily at 21:00 UTC, and the first 24h of prediction with 3-hourly outputs are used for training. The fields considered are the two horizontal directions of wind at 10-meter height, referred to as u (zonal component) and v (meridian compo-

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

nent), and the 2-meter temperature (t_{2m}). Additionally, as will be detailed in Section 5, orography is used in some experiments as a constant field for the GAN to generate.

In order to provide a flexible experimental setup, and to keep the training runs in reasonable time windows, a small sub-region of the AROME domain is selected. It corresponds to the Mediterranean coastal region and the Rhône valley (see Figure 1). The sub-domain considered thus spans 128×128 grid points, with an approximate 330 km side size. This choice of localisation is motivated by the variable terrain features and identifiable weather patterns known to occur in this region. The joint presence of the French Alps and Mediterranean sea, as well as marked episodes of strong northerly winds (mistral events) and heavy precipitating events characterized by localized strong gradients, both made an interesting case for trials on this region. Moreover, the high resolution of the samples allows for investigation of several scales of variability, from regional scale down to the typical grid scale of state-of-the-art convection-permitting models.

In the baseline configuration, fields of u, v, t_{2m} for a given lead time, date and ensemble member are learned jointly as part of the same sample. One single day of forecasts hence yields $8 \times 16 = 128$ distinct data samples. Altogether, the usable dataset is then composed of 66048 samples. The shape of the GAN output tensors is then $3 \times 128 \times 128$, with 3 being the number of variables and 128×128 the domain size.

Using a dataset from AROME-EPS increases the volume of training data, compared to using the deterministic AROME forecasts over the same period. It is important to note that many of the samples are indeed correlated, whether they correspond to close lead times or to different members of a same forecast. However, for a given forecast, samples corresponding to different lead times and different ensemble members are physically distinct, namely because of the fine-scale variability of wind fields, and of the diurnal cycle of temperature. The ensemble-based dataset thus exhibits an increased small-scale diversity compared to a deterministic dataset. It is thus possible to view the ensemble as a NWP-specific data augmentation strategy, implemented upon a deterministic forecast system on a given period. This study will thus assess to what extent a GAN can recover this given, fixed distribution of high resolution samples enriched by the EPS.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

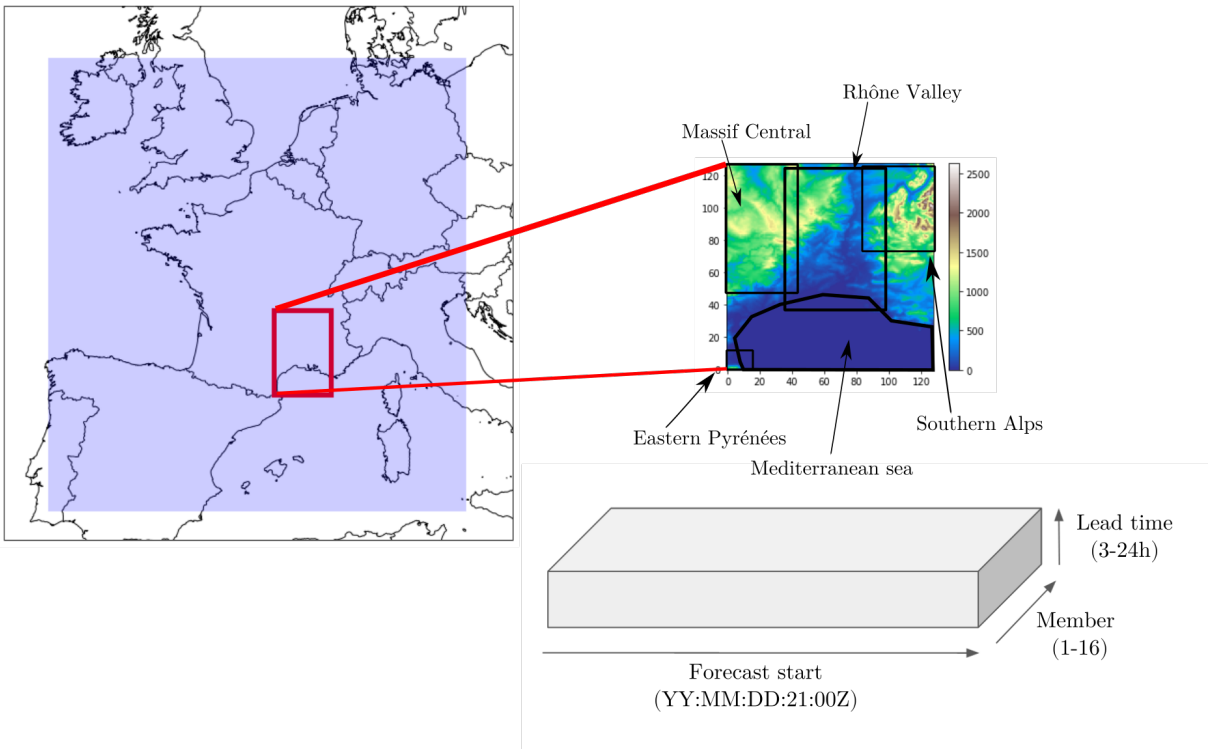

FIG. 1. On the left, the full AROME domain is shown, along with the 128×128 sub-domain used for the GAN training. Its main geographical features are represented on a topographic map (top right, with altitude in meters). Dataset organisation with its main variability directions is shown on the bottom right. Each sample is a $3 \times 128 \times 128$ array, corresponding to a "volume element" of the "dataset box" shown at the bottom right.

b. Choices for the GAN architecture

The GAN framework has been thoroughly investigated in recent years, and some guidelines have emerged to design efficient and reliable GAN training algorithms. Let us denote with $\mathbb{P}_{data}$ the distribution of the target dataset to be emulated (in our case, AROME-EPS). The purpose of the GAN is to provide a function G mapping a high-dimensional, *a priori* defined distribution $\mathbb{P}_z$ onto $\mathbb{P}_{data}$. $\mathbb{P}_z$ is defined on a latent space Z taken as input to the deep network supporting G (the generator). Outputs from G are then given as inputs to the discriminator network D, which tries to distinguish between "fake" samples outputted by $G(Z)$ and the "real" samples $X \sim \mathbb{P}_{data}$. The output of D is then a single scalar assigning a "score" to each sample.

The objective function must ensure that outputs from D confer high scores to the "real" distribution while maintaining low scores on outputs from G. D being optimized, G then aims at

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0006.1.

producing samples *confusing* D, i.e obtaining high scores from D. An optimal training then results in D being unable to separate "real" samples from "fake" ones, though being optimally designed to separate them. Ideally then, the distribution produced by G completely and correctly recovers $\mathbb{P}_{data}$.

The GAN training framework exhibits convergence and stability issues. Notably, the well-known "mode collapse" problem consists in G concentrating the mass of $\mathbb{P}(G(Z))$ on a small part of the $\mathbb{P}_{data}$ distribution while "forgetting" about the rest. To tackle this phenomenon, Arjovsky et al. (2017) emphasize the need for the discriminator to be a smooth (Lipschitz) function so that it continuously separates "fake" and "real" distribution samples, and introduce the Wasserstein-GAN framework (WGAN), noting that the discriminator is trained to assess a Wasserstein distance between the "fake" and "real" distributions. Making the discriminator a Lipschitzian function of its input samples comes down to bounding its gradient. Several formulations of the GAN's objective have since then implemented this regularization constraint, which effectively improved on the original GAN formulation. The guidelines of Miyato et al. (2018), i.e. use spectrally-normalized convolution layers, are used in this work. Spectral normalization (SN) consists in renormalizing the weight matrices of the discriminator to bring their highest singular value to one (hence the term 'spectral'). SN thus naturally enforces the Lipschitz constraint while being more efficient than other techniques such as gradient clipping (Arjovsky et al. 2017), which imposes an arbitrary upper boundary on the gradient, or gradient penalty (GP, Gulrajani et al. (2017)), which penalizes the gradient when its norm deviates from unity.

To the author's knowledge, this study is one of the first to propose a WGAN-SN framework for geophysics applications. Among the studies dealing with the generation of atmospheric fields with GAN, only Ravuri et al. (2021) use SN, while others, such as Besombes et al. (2021) and Harris et al. (2022), keep using the WGAN-GP formulation. Miyato et al. (2018) only apply spectral normalization on discriminator layers, but literature since then has acknowledged the positive effect of using SN on both generator and discriminator (Brock et al. 2018). This double regularization is implemented in the present setup.

GANs can also produce unstructured or strongly corrupted samples, depending on what features from the dataset the discriminator is able to identify as crucial. This failure mode can happen in the absence of mode collapse, or concomitantly to it (Brock et al. 2018; Mescheder et al. 2018). This

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

makes the training of GANs difficult, even within the Wasserstein-GAN framework, and special care has to be devoted to the networks' hyperparameters choice.

Finally, the "hinge-loss" objective formulation given by Lim and Ye (2017) is used, where both G and D are trained to minimize their loss:

$$\begin{aligned}\mathcal{L}(D) &= \mathbb{E}_{X\sim\mathbb{P}_{data}}\left[\max(0, 1 - D(X))\right] \\ &\quad + \mathbb{E}_{Z\sim\mathbb{P}_z}\left[\max\left(0, 1 + D(G(Z))\right)\right] \\ \mathcal{L}(G) &= -\mathbb{E}_{Z\sim\mathbb{P}_z}\left[D(G(Z))\right]\end{aligned}$$

These quantities are minimized through stochastic gradient descent, using an Adam Optimizer (Kingma and Ba 2015) to adapt the weights of the two networks. Practically, expectations are estimated by randomly drawing samples from the AROME-EPS dataset (for the $\mathbb{P}_{data}$ estimator) and from the $\mathbb{P}_z$ distribution. The parameters of D and G are then updated alternatively. We set $\mathbb{P}_z$ to a centered, normal distribution of dimension $d = 64$ with identity covariance matrix. The choice of d was determined by following previous literature (Besombes et al. 2021; Mustafa et al. 2019). This dimension was fixed all along the study, since Marin et al. (2021) indicate that this parameter might not have a significant influence in the quality of generation, provided it is not too small (typically, below 64).

A residual architecture is chosen, consisting of two residual nets directly taken from Miyato et al. (2018), as shown in Figure 2. Starting from a 64-dimensional latent space, samples are shaped into feature maps whose resolution gradually increases as they go throughout the generator, to be finally outputted through a *tanh* layer. The discriminator follows a symmetric pattern downscale, though being one layer deeper and involving more channels before the last dense, output layer. It should be emphasized that, although the networks are relatively shallow, their estimated receptive fields (i.e the size of the input region which contributes to a given output pixel) are large enough to model long-range correlations. For example, a single residual block of the generator involves two 3×3 convolutions and one upsampling layer with a dilation factor of 2. This configuration gives a maximal local receptive field of 10 pixels for each residual block. Since we stack 5 of these blocks, the final, global receptive field spans beyond 128 pixels, which is the size of our samples.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

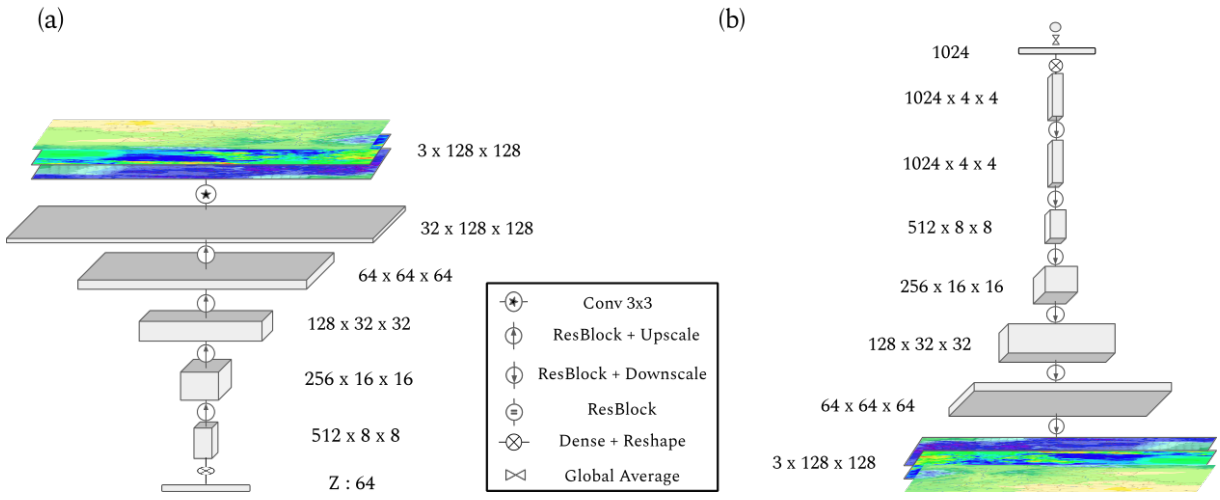

FIG. 2. Network architectures of generator (a) and discriminator (b). Input layers are at the bottom of the schematic and networks process tensors from bottom to top.

Therefore, the final sample is able to take into account each degree of freedom from the random input.

Before training, the samples are re-scaled so that the global minimum and maximum values of the dataset fit within the $[-0.95, 0.95]$ range. This is to ensure that the hyperbolic tangent output of the generator can reach the dataset's extremes and even go beyond these limits. The mean, minimum, and maximum value are pre-computed over all grid points and all data samples. Supplementary training parameters and procedures (floating precision, warm-up, initialization) are detailed in Appendix A. These models are trained for a fixed number of $60,000$ update steps, on a cluster of 4 NVIDIA V100 GPUs with 32GB of RAM. Models are thus trained for 4 to 12h wall-clock time, depending on the batch size. A step is equivalent to one update of both networks after forward and backward pass. Thus, for two different batch sizes, this fixed number of steps allows for different numbers of epochs (i.e sets of steps corresponding to 100% of the dataset seen by both networks).

3. Evaluation metrics

a. Distributional metrics

The training is monitored thanks to three estimates of earth-mover distance (EMD, Rubner et al. (2004)) between $\mathbb{P}_{data}$ and $\mathbb{P}_{GAN} = G(\mathbb{P}_z)$. This metric allows for quantification of the proximity of

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0006.1.

two multidimensional distributions. The exact retrieval of the EMD is a hard problem in general, and approximate EMD estimators converge slowly, necessitating large number of samples to be accurate (Ramdas et al. 2015). However, when the data is univariate, one is left with:

$$W_1(\mathbb{P}, \mathbb{Q}) = \int_0^1 |F_P^{-1}(t) - F_Q^{-1}(t)| dt$$

Where F_P, F_Q are the cumulative distribution functions associated to $\mathbb{P}$ and $\mathbb{Q}$ respectively.

Following the strategy of Besombes et al. (2021), two 1D-EMD estimates are computed: at each test step, a random number of pixels is sampled, to evaluate the average of per-pixel, per-variable 1D-EMDs ; another average of 1D-EMDs is also evaluated on a fixed number of pixels, covering the central 64×64 crop of the domain (and averaged over variables). These estimates are hereafter termed $W_{1,\mathrm{r}}$ (random pixels) and $W_{1,\mathrm{c}}$ (central crop), and are a global measure of the quality of the generation of marginal (per-pixel, per-variable) distributions.

A third estimate of EMD is taken from Karras et al. (2018) and termed *multi-level sliced Wasserstein distance* (SWD_{multi}). This estimate measures multi-dimensional EMDs of images at different resolutions. Precisely, it decomposes the image signals on 4 different resolution levels and generates a Laplacian pyramid: starting from the finest-grained level, each image is obtained from the previous by Gaussian filter convolution, difference, and subsampling. One then measures EMD on each of the levels obtained, for the joint distribution of all variables. These estimates go from the fine-grained level (where images have 128×128 dimensions and conserve small-scale fluctuations) to the coarse-grained level (where images only have 16×16 dimensions and conserve low frequency fluctuations). They are used to compare the distribution of patterns at each level. The name "*sliced Wasserstein distance*" (SWD) refers to the way the estimates of EMD on multi-dimensional spaces are performed. This estimation procedure is unbiased (Rabin et al. 2011; Kolouri et al. 2018), and robust as long as the number of samples is large enough. This yields a 4-component distance (one component for each level): SWD_{multi} = (SWD_{128}, SWD_{64}, SWD_{32}, SWD_{16}). Following the intuition of Rabin et al. (2011), the SWD_{multi} metric follows a kind of 'wavelet decomposition' approach, as it measures the discrepancy between two image distributions at several scales of fluctuations, for local neighbourhoods, merging the contributions of different variables. A visual explanation of this metric is shown in Figure 3 and a detailed description is given in Appendix B.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

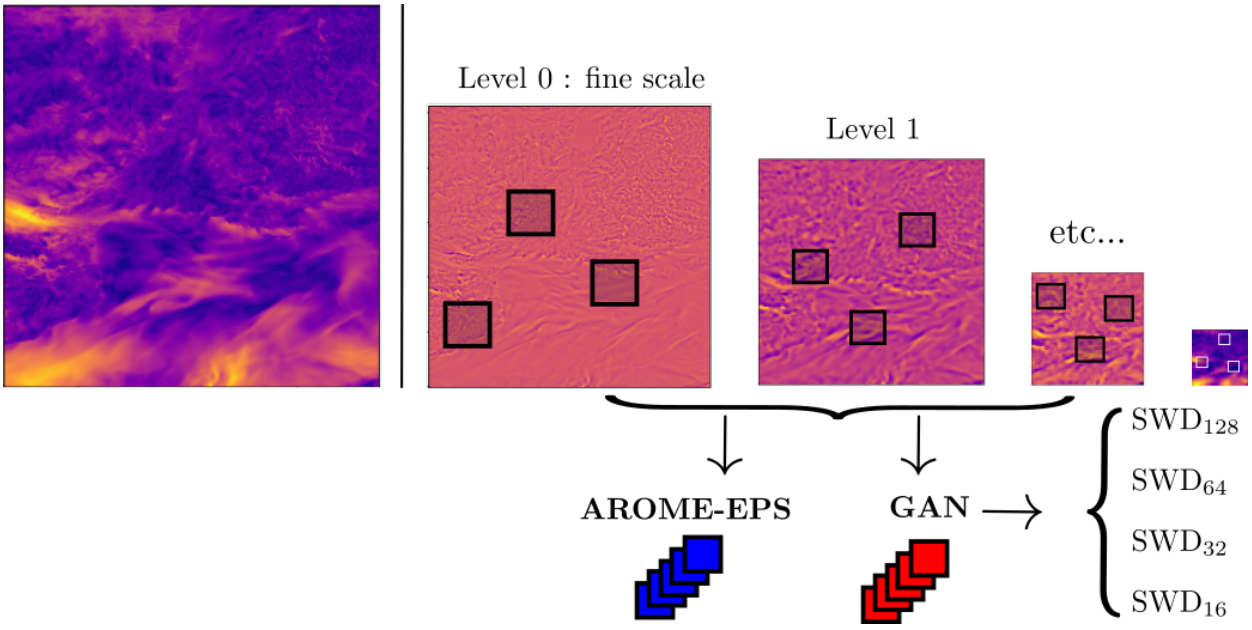

FIG. 3. Schematics of SWD_{multi} computation. Starting from a base image (left), several levels are recursively created via a Laplacian pyramid procedure. Random neighbourhoods are then selected (small black squares) for each level, allowing for estimation of the EMD on these neighbourhoods for each level, thanks to the SWD algorithm. Appendix B details this procedure.

A lower bound for these distances is set by their estimate from the AROME-EPS to itself. While the theoretical distance is 0, estimating it through finite batches from the AROME-EPS dataset yields a positive result. This distance is estimated by a bootstrap procedure. We select several batches of 16384 samples (with no replacement within one batch, but with replacement from one batch to the other), and then compute the EMD estimates between pairs of these batches. The average value of the pair-EMD series is kept as the distance estimate (see Table 1). 16384 samples represent nearly 25% of the full dataset: we thus deem that this procedure represents correctly the diversity of the dataset. The values obtained correspond to lower bounds for EMDs, as they reflect the internal variability of the dataset and account for finite sampling effects. If an EMD estimate reaches the lower-bound values, it can be said that the GAN dataset would be completely indistinguishable from the AROME-EPS dataset regarding this estimate.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0006.1.

Metric	$W_{1,r}$	$W_{1,c}$	SWD_{128}	SWD_{64}	SWD_{32}	SWD_{16}	SWD_{avg}
Score ($\times 10^{-3}$)	1.4/0.3	1.4/0.3	1.5/0.1	1.5/0.1	1.6/0.1	4.6/0.9	2.3/0.2

TABLE 1. Estimates of the EMD from the AROME-EPS dataset to itself. These are estimated from 32 independent selections of 16384-batches pairs. Shown is average/standard deviation for the series of tested pairs.

b. Power Spectral Density Error

The EMD estimates are completed with the average power spectral density (PSD) spectrogram obtained from AROME-EPS and GAN samples, for each variable. A root mean-square error (RMSE) is taken on the difference of spectrograms (logarithmic scale) to give each scale the same weight. It is measured in decibels (dB) and reads:

$$\begin{aligned} PSD_{err} &= \sqrt{\left\langle \left(10\log[PSD_{GAN}] - 10\log[PSD_{AROME}]\right)^2 \right\rangle} \\ &\approx 10\sqrt{\left\langle \left(\frac{PSD_{GAN}}{PSD_{AROME}} - 1.0\right)^2 \right\rangle} \end{aligned}$$

Where $\langle\cdot\rangle$ denotes average over spectral scales. This compares the deviation of the GAN from the AROME spectral repartition of energy. The PSD_{err} metric is frequently used to assess the realism of GAN predictions (Leinonen et al. 2021; Ravuri et al. 2021; Harris et al. 2022), where errors of a few dB are usually considered as good quality. Spectrograms are computed with discrete cosine transform (Denis et al. 2002) to avoid aliasing effects due to the non-periodicity of our samples. While this metric provides a sound evaluation of sample quality, it does not provide a complete view of the organization of bi-dimensional fields, or of multi-scale interactions. Namely, Gaussian noise fields can be tuned to recover the exact 2D spectrum of any other, fixed 2D field (Bruna and Mallat 2013). On the other hand, evaluating the diversity of samples produced by the GAN requires metrics that can evaluate the proximity between probability distributions, rather than between individual samples.

Going further requires to consider supplementary diagnostics in the evaluation procedure, which are presented in the remainder of this section.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

c. Correlation lengths

The PSD metric has a known drawback: as a completely non-local metric, it cannot provide insight into the spatial distribution of field's variability. In other words, it is not sensitive to complex, hierarchical textures. Accounting for local structures and cross-scale interactions is easier with metrics involving calculation of local quantities. An example is the local correlation length scales. The correlation length scale can be defined as the typical length scale over which a given field is spatially correlated to itself. Large correlation lengths on a given grid point indicate that this point is often part of long-range structures. This diagnostic is common in data assimilation (Pannekoucke et al. 2008; Raynaud and Pannekoucke 2013), where it can be used to fit the error covariance matrix. Correlation lengths are also related to semi-variogram scores (Olea 1994), commonly used to evaluate the performance of meteorological models. According to Weaver and Mirouze (2013), this length scale can be given as:

$$L_{corr} = \sqrt{\frac{-2}{\mathrm{Tr}[g_{xy}]}}$$

Where $g_{xy} = -\mathbb{E}\left[\partial_x X \partial_y X\right]$ is the mean local metric tensor obtained from the normalized field X and $\mathrm{Tr}[\cdot]$ represents the trace operator. The $\partial_x X$ (resp. $\partial_y X$) notation corresponds to the spatial gradient of X in the zonal (resp. meridian) direction. These gradients are estimated from the grid data using finite differences ; the expectation of $\partial_x X \partial_y X$ is then taken over the samples of a given dataset. The resulting g_{xy} tensor can be evaluated on each grid point and for each variable, yielding maps of L_{corr} for each variable. The trace operator makes L_{corr} an isotropic quantity ; further manipulation of the g_{xy} tensor can provide insights about local anisotropy (Pannekoucke et al. 2008) but we chose to keep the simpler L_{corr} quantity as an indicator of structure sizes.

d. Second-order scattering coefficients

Here we present Scattering Coefficient approaches (Andén and Mallat 2014; Bruna and Mallat 2013; Cheng et al. 2020), and show how they can be used to extract useful information from our meteorological dataset. Scattering Coefficients are derived from recursive wavelet filtering of the fields. They have already been applied to meteorological information classification with satisfactory results, e.g in Garcia et al. (2015). The calculation of first and second-order scattering

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

coefficients is detailed in Appendix B. Let λ_1 and λ_2 denote scales with $\lambda_1 < \lambda_2$. The scales considered can go from 2 grid points (circa 2.5 km) to 64 grid points (circa 160 km), and wavelet filters ψ for first and second-order coefficients have different orientations θ_1, θ_2.

The convolution of a field X with $\psi_{\lambda_1,\theta_1}$, followed by modulus, yields a set of *first-order* scattering maps $M_1(\lambda_1)$ and resulting first-order coefficients:

$$M_1(\lambda_1, \theta_1) = |X \star \psi_{\lambda_1,\theta_1}|, \quad S_1(\lambda_1) = \langle M_1(\lambda_1, \theta_1) \rangle$$

where $\langle \cdot \rangle$ denotes spatial averaging and dependence to θ_1 is omitted.

First-order scattering coefficients $S_1(\lambda_1)$ relate to the intensity of signal energy at scale λ_1. As indicated by Cheng et al. (2020), they play a similar role to spectral decomposition.

Convolving again with $\psi_{\lambda_2,\theta_2}$ and taking modulus yields second-order maps and coefficients:

$$M_2(\lambda_1, \lambda_2, \theta_1, \theta_2) = |M_1(\lambda_1, \theta_1) \star \psi_{\lambda_2,\theta_2}|$$
$$S_2(\lambda_1, \lambda_2) = \langle M_2(\lambda_1, \lambda_2, \theta_1, \theta_2) \rangle$$

Second-order coefficients $S_2(\lambda_1, \lambda_2)$ with $\lambda_2 > \lambda_1$ measure the energy at λ_2 of a signal which has already been filtered to show variations of scale λ_1 only. This accounts for the organisation of patterns of typical scale λ_1 on a larger scale λ_2. They give an average effect of multi-scale interactions.

Figure 4 shows the chain of transforms: starting from a full AROME image, successive scattering steps are applied. Regions with sharp wind gradients are highlighted by the first pass of wavelet at scale $\lambda_1 = 2$ pixels. The second-order pass enhances clusters of such regions at scale $\lambda_2 = 4$ pixels. Such clusters contribute significantly to the S_2 coefficient (e.g, the top right and center left green frames). Regions with sharp variations organisation at larger scales (central green frame) contribute to a lesser extent, while quasi-uniform regions are smoothed out, and have only negligible contribution (white, bottom left frame).

Scattering coefficients can be used to measure the "sparsity" of a signal. A sparse signal will concentrate small-scale variations on localized points, and exhibit large-scale organisation of

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

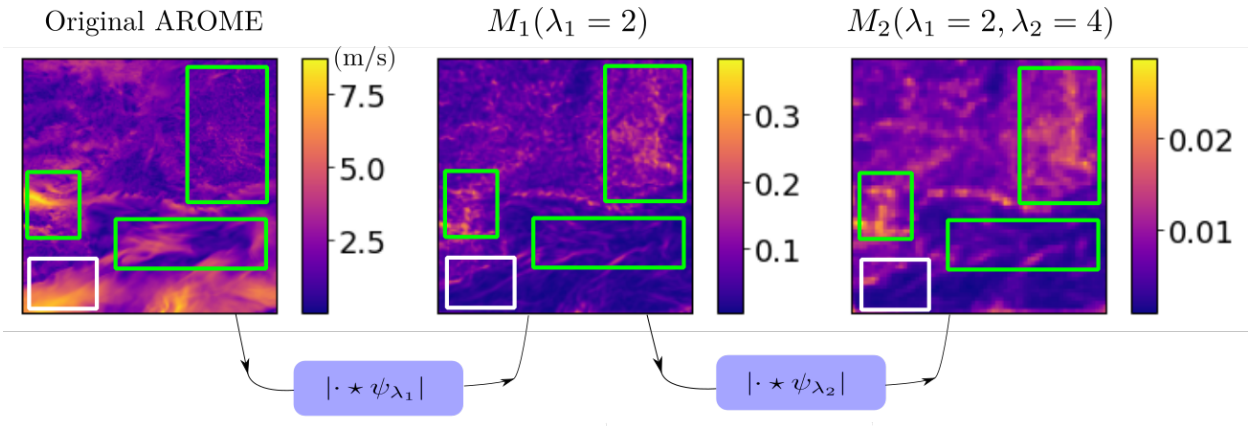

FIG. 4. Transformation chain for scattering coefficients. The base (left) image represents wind speed for an AROME-EPS field. The center image represents a first-order, small-scale scattering map, and the right image is a second-order, medium scale scattering map. Regions of interest are framed: regions having an important S_2 contribution in green, and a region with a negligible S_2 contribution in white. Spatial subsampling is due to the wavelet decomposition algorithm (Andreux et al. (2018)).

these local variations. Cheng and Ménard (2021) introduced the following quantity as a sparsity estimator:

$$s_{21}(\lambda_1, \lambda_2) = \left\langle \frac{S_2(\lambda_1, \lambda_2)}{S_1(\lambda_1)} \right\rangle_{\theta_1, \theta_2}$$

A high value of s_{21} for a given scale pair λ_1, λ_2 indicates that a significant quantity of signal information lies in the second-order coefficient: the organisation of the λ_1 map is important to account for the image global structure, and the field is rather "sparse". On the contrary, a field exhibiting a uniform λ_1 map would have a low s_{21} for $\lambda_2 > \lambda_1$ (since only few regions would exhibit variations at scale λ_2). Typically, Gaussian noise maps are not "sparse": most of the information they contain can be extracted from their spectrum, or almost equivalently from their S_1 coefficients. In Appendix B we show that a Gaussian noise field with the exact spectrum of AROME-EPS samples exhibits lower s_{21} than AROME-EPS samples.

Averaging over orientation reduces the number of coefficients at the expense of orientation-related information. To obtain information about fields anisotropy, Cheng and Ménard (2021) proposed a "shape" estimator:

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0006.1.

$$s_{22}(\lambda_1, \lambda_2) = \left\langle \frac{S_2(\lambda_1, \lambda_2)_{\theta_1=\theta_2}}{S_2(\lambda_1, \lambda_2)_{\theta_1 \perp \theta_2}} \right\rangle_{\theta_1}$$

The s_{22} estimator helps determining in what directions multi-scale interaction is most likely to happen, regardless of the initial orientation of patterns (represented by θ_1). They interpret large s_{22} as a marker for filaments (with $\theta_2 = \theta_1$ directions producing higher S_2 values), while lower s_{22} indicates the presence of more "roundish" shapes (with $\theta_1 \perp \theta_2$ directions producing higher values). Again, this measure goes beyond radially-averaged spectra, as a Gaussian field with the spectrum of AROME-EPS shows lower s_{22}.

Both s_{21} and s_{22} estimators provide a set of coefficients (one for each λ_1, λ_2 couple with $\lambda_1 < \lambda_2$). One can estimate the discrepancy between the GAN and AROME-EPS data, for both the s_{21} and s_{22} distributions, as a measure of good reproduction of atmospheric structures. The sets of average s_{21} and s_{22} can be used to calculate RMSE distances between AROME-EPS and the GAN, yielding two new metrics per variable. These distances are evaluated during training with 16384 batches, and serve as quality estimators, similarly to PSD errors.

e. Complementary metrics

Cross-variable correlations will be assessed using bi-dimensional histograms, similarly to those used by Gagne II et al. (2020). For a given pair of variables and a given dataset, these histograms provide the empirical density function for the values taken by the pair of variables. Ideally, the densities outputted by the GAN should overlap the densities extracted from the AROME-EPS. This is a graphical illustration of the precision-recall metrics used, e.g., in Kynkäänniemi et al. (2019). Among others, this enables the identification of systematic biases in the GAN-produced distribution.

Finally, maps of 10^{th} and 90^{th} percentiles and inter-percentile range will be examined, in order to focus on the representation of distribution tails.

The whole set of metrics used is summarized in Table 2, detailing attributes for each. As explained above, the metrics are used to measure either diversity or quality. They make use of several types of information: *pixel-wise* information does not take any spatial correlation into account, *neighbourhood* information represents the aggregation of information over a limited range of pixels, and finally *non-local* information aggregates information over the full sample scale, either

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

by Fourier transform, or random sampling of neighbourhoods. Finally, metrics measure different features of the signal, notably *scale-by-scale* information, *multi-scale* organisation, *positional* information (when the metric is plotted on a map) or *anisotropy*.

	Metric	$W_{1,r/c}$	SWD_{multi}	2D histograms	Quantiles	PSD_{err}	s_{21}	s_{22}	L_{corr}
	Definition	EMD pixels	EMD patterns	Cross-var corr.		Spectral err.	Sparsity est.	Shape est.	Corr. length
Purpose	Diversity	✓	✓	✓	✓	×	×	×	×
	Avg. Quality	×	×	×	×	✓	✓	✓	✓
Inform. used	Pixel-wise	✓	×	✓	✓	×	×	×	×
	Neighbourhood	×	✓	×	×	×	✓	✓	✓
	Non-local	×	✓	✓	×	✓	✓	✓	×
Measures	Per-scale info.	×	✓	×	×	✓	×	×	×
	Multi-scale org.	×	×	×	×	×	✓	✓	×
	Positional info.	✓	×	×	✓	×	×	×	✓
	Anisotropy	×	×	×	×	×	×	✓	×

TABLE 2. Summary of the metrics used in the text. Checkmarks indicate which attributes a given metric possesses, crosses indicate the absence of such attribute.

f. Evaluation strategy

We chose not to *a priori* separate the dataset between training, validation and testing data. Though it is similar to the methodology of Besombes et al. (2021), this is arguably not a common practice in machine learning. It nevertheless makes sense in our setup, for the following reasons:

- The generator is unconditional and never takes any other inputs than latent vectors. Its ability to generate good-quality, high-resolution samples and a correct distribution *once trained* only depends on the mapping it makes between the Gaussian latent distribution and the distribution of samples in the 'physical space'.

- The aim of this study is to assess what features of the data distribution the GAN is able to produce. Comparing the output distribution of the GAN to the *training* distribution thus avoids to take into account the necessary distribution shift that occurs when splitting the dataset between train, test and validation.

- Detecting mode-collapse, or more generally the loss of diversity, can be done through the combined use of EMD distances, since these metrics provide large scores to distributions with

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

too low spread. Cross-variable correlations can also help detecting biased or non-overlapping distributions.

- Loss of quality can be examined through average PSD error, average correlation length scales estimation and average scattering coefficient errors, each metric having its own scope.

To compare different hyperparameter sets, we focus exclusively on EMD distances and PSD error. Once a satisfactory set is selected, we examine the samples produced by the GAN with the other metrics.

To be consistent with the lower bound estimation procedures detailed in Subsection 3a, each metric is applied to 16384 random samples from the AROME-EPS dataset and to the same number of random GAN outputs. Especially for SWD_{multi} distances, this rather large number of samples reduces the estimator's variance, as it done in Odena et al. (2017) and Karras et al. (2018).

4. Results

a. Training stability and convergence

Even with widely used regularization strategies, training a GAN requires careful tuning of parameters to ensure local convergence.

Initializing with a learning rate $lr_0 = 4 \times 10^{-3}$ and using exponential learning-rate decays ($lr = lr_0 \cdot \gamma^t$ with t the number of epochs and $\gamma = 0.9$) avoids mode collapse and produces realistic-looking samples for all batch sizes except the largest (512). Setting different learning rates or different decay rates γ for D and G was detrimental to quality and convergence, even from a mere visual perspective. In agreement with Mescheder et al. (2018), removing learning rate decay had terrible results on performance, leading to severe mode collapse and forcing each pixel to the global dataset average value (close to 0). Keeping $\gamma = 0.9$, we select the best performing configuration among several batch sizes and learning rates according to estimates of $W_{1,r/c}$, SWD_{multi} and PSD errors, keeping the rest of metrics for post-training evaluation. The configuration with a batch size (BS) of 32 and $lr_0 = 4 \times 10^{-3}$ is selected, as it globally has the lowest distributional distances and PSD errors simultaneously. Details of the hyperparameters selection are shown in Appendix C.

Table 3 shows the scores obtained at the end of the run (i.e. when the loss curve plateaus). Figure 5 compares the average spectrum produced by the GAN to the AROME-EPS spectrum, showing

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

Metric	$W_{1,r}$	$W_{1,c}$	PSD_u	PSD_v	$\mathrm{PSD}_{t_{2m}}$	SWD_{128}	SWD_{64}	SWD_{32}	SWD_{16}	$s_{21,u}$	$s_{21,v}$	$s_{21,t_{2m}}$	$s_{22,u}$	$s_{22,v}$	$s_{22,t_{2m}}$
Unit/Scale	$\times 10^{-3}$		dB			$\times 10^{-3}$				$\times 10^{-3}$			$\times 10^{-2}$		
Score	13	12	8.1	8.9	11	5.7	7.3	12	39	5.0	3.8	5.6	4.7	4.6	1.7

TABLE 3. Scores obtained with the different metrics used by the best-performing configuration of hyper-parameters: PSD errors, SWD estimates, and RMSE of Scattering estimators with respect to AROME. Each configuration was run 3 times to account for training variability ; the scores presented are the best obtained among these runs. Appendix B (especially Figures B1 and B2) provides means to interpret the absolute values hereby provided.

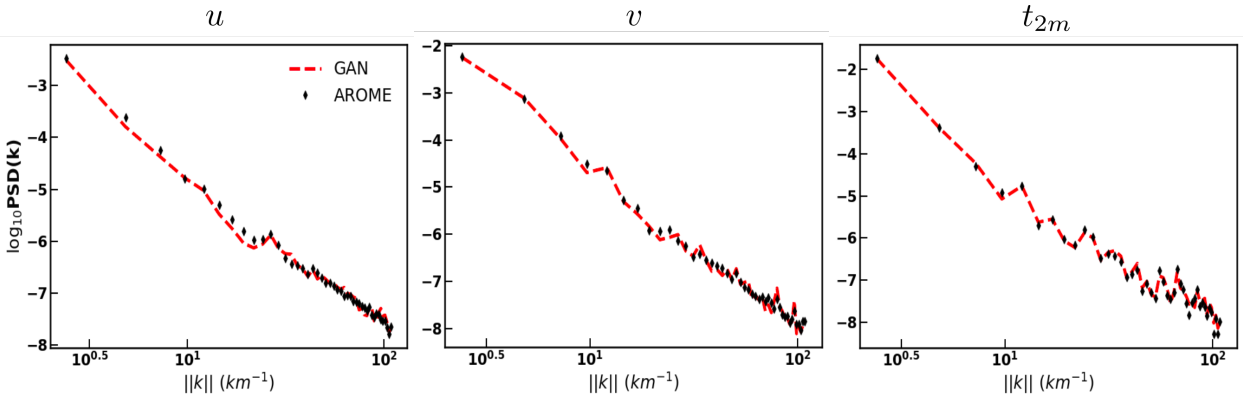

FIG. 5. Typical GAN PSD spectrograms (red line) obtained for the best configuration ($BS = 32$, $lr_0 = 4 \times 10^{-3}$), for each variable. Black dots indicate the average AROME-EPS spectrum.

good agreement and low error (below or around 1 dB for each variable), for all scales. Even the sharp variations of temperature spectra (mostly due to topographic variations) are correctly reconstructed despite slightly higher PSD errors. Notably, for the smallest scales, no significant drop of PSD can be observed, meaning that even large wave numbers are given a correct amount of energy.

One noticeable feature of Table 3 is the large difference observed from one SWD component to another: they improve as the component's scale decreases. Such behaviour was observed with several hyperparameter configurations. Small-scale components such as SWD_{128} and SWD_{64} get the lowest (i.e best) scores: small-scale distributions of local patterns are hence more correctly fitted, at least by the best configurations. On the other hand, even these configurations have sensibly higher SWD_{32} and SWD_{16} values. The floor values given in Table 1 are indeed larger for SWD_{16} than for other SWD components, indicating a larger intrinsic variability of the dataset for these

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0006.1.

scales. However, the gap between the scores of the GAN capabilities and the AROME lower bound is larger for SWD_{32} and SWD_{16}: this observation might then indicate a poorer fit of large-scale pattern diversity. This result was observed for all hyperparameter configurations (cf. Appendix C). It is then reasonable to think this is related to the network architecture or to the training set-up.

In the remainder of this section, we provide an in-depth analysis of the GAN performances using the set of metrics previously introduced.

b. Validation of results

1) Visual examination

As a first step, the quality of the GAN generations can be assessed subjectively. Some samples are presented on Figures 6 and 7 for visual comparison.

First, the resolution of GAN-produced samples appears to be correct when compared to AROME-EPS outputs. As can be seen especially on temperature maps, fine-grained mountainous regions are correctly generated by the GAN, with a visible cooling with altitude. More detailed comparisons can be made with the help of the geographical features given on Figure 1. Lengthy wind structures are preferentially created over sea, while they are more granular over land. Specific weather patterns are also present in the GAN's samples, such as strong northerly wind running down the Rhone Valley, an event locally known as 'mistral'. Figure 7 also shows that the GAN is capable of producing consistent wind direction and speed at the highest detail level. The wind map especially confirms the ability of the GAN to generate not only events such as mistral, but also a rather large diversity of wind patterns. Qualitatively speaking, our GAN is thus arguably devoid of mode-collapse. What is more, Appendix D shows that the GAN does not simply memorize exact samples from the dataset but indeed produces unseen, distinct data samples.

On the other hand, the GAN often produces blurry wind patterns over sea, while AROME-EPS samples are significantly more structured. The clear wind fronts present in the AROME-EPS dataset also appear in the GAN's samples, but, except for the mistral case, they lack a long-range consistency and the 'filamentary' aspect of AROME-EPS wind.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

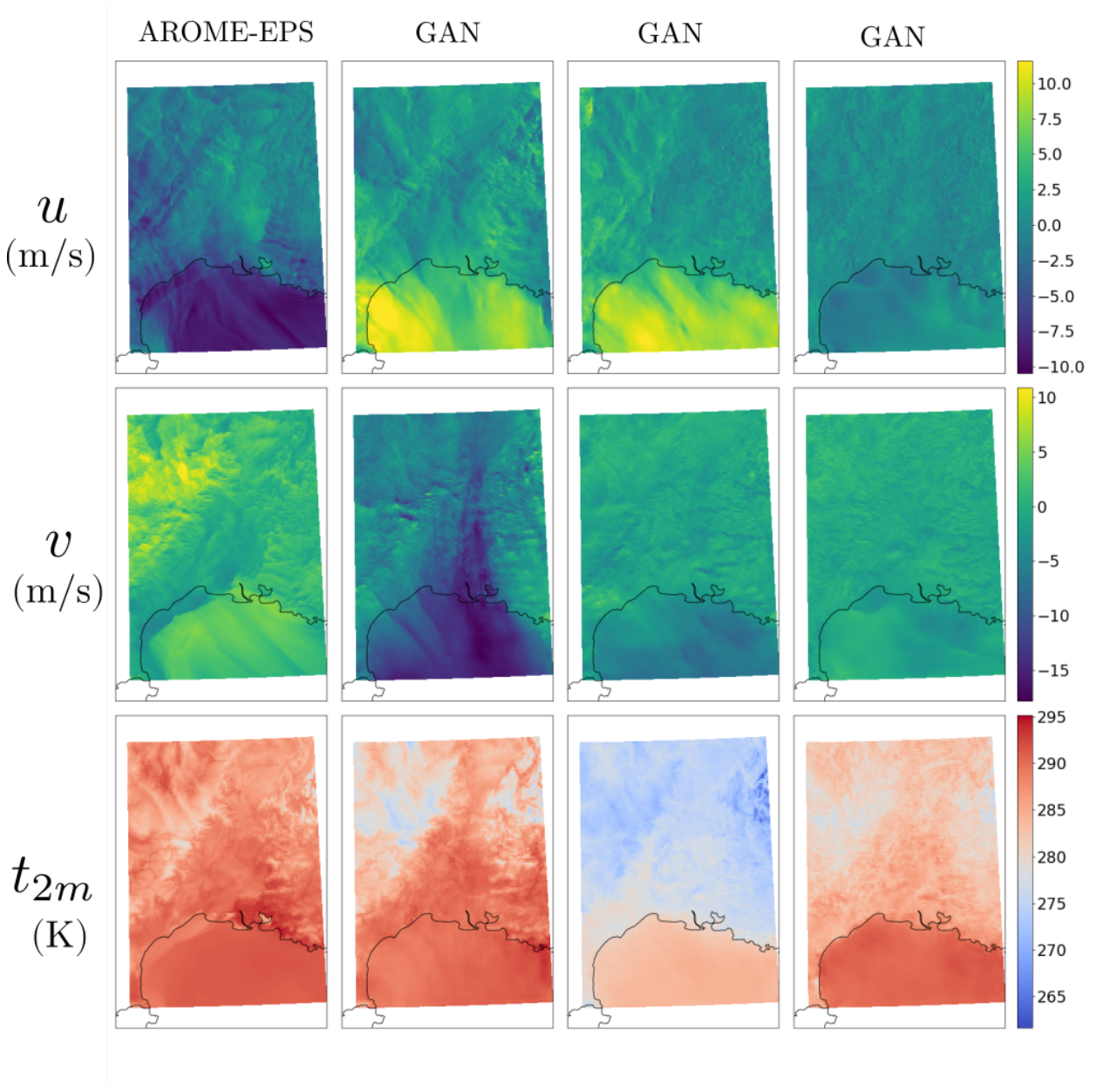

FIG. 6. Left column: a random AROME-EPS sample. Right columns: random samples from the GAN. The AROME-EPS samples are provided for visual comparison only, as there is no one-to-one correspondence between the AROME-EPS samples and the GAN samples.

2) EMD MAPS

Pixel-wise maps of EMD for chosen iterations indicate which regions provide the largest divergence. Figure 8 shows a comparison between the original dataset variance for each variable, and

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0006.1.

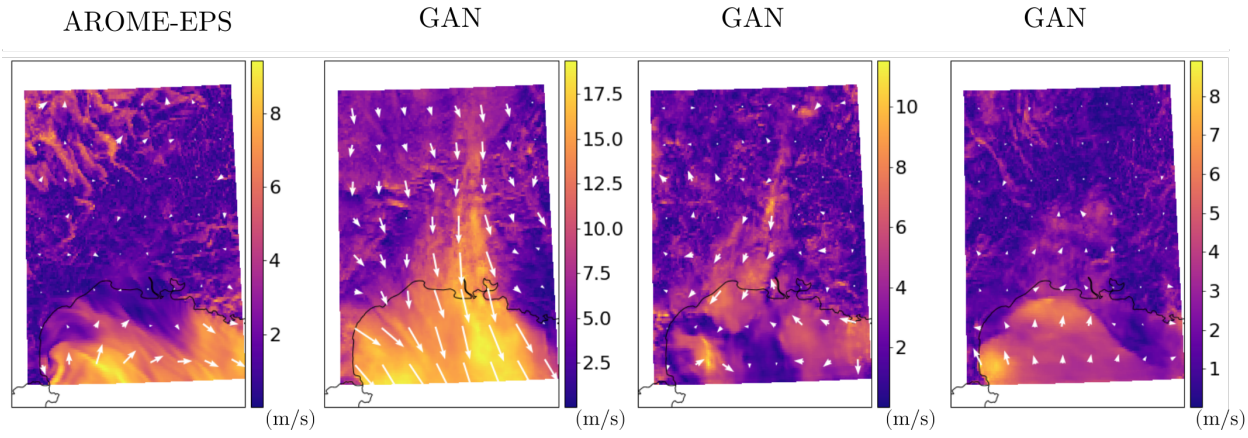

FIG. 7. Wind speed and superimposed wind direction. From left to right: AROME-EPS, random samples from the GAN. Arrows are regularly spaced, with length proportional to intensity. The GAN produces diverse and consistent situations for the wind, but lacks the detailed structure of the AROME data, especially over sea.

two maps of pixel-wise EMD at two different steps of training. Regions with highest variance are globally less easily learnt by the GAN than their low-variance counterparts, especially at the beginning of training. The land-sea mask is well visible here, as can be expected from physical arguments. Indeed, wind variability over sea is higher: it depends more on the global weather situation (e.g., presence of fronts) and is not forced by the topography. It coincides with long-range structures with a broad range of directions and intensities. This is not the case over land, where surface plays a major role in reducing the range of correlations. The northerly mistral path is clearly highlighted, as well as the easternmost and westernmost wind variability poles (roughly corresponding to 'tramontane' wind episodes in the west). On the other hand, sea temperature is relatively stable because of the water thermal inertia while the diurnal cycle is far more pronounced over land. Especially, temperatures of mountainous summits are difficult to reproduce, because cold extremes of the distribution are susceptible to occur there.

This implies variance-related error is probably a strong learning signal for the GAN, especially at the beginning of the process. It is consistent with the observations made in the previous subsection about position-related distributions being the easiest to fit.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

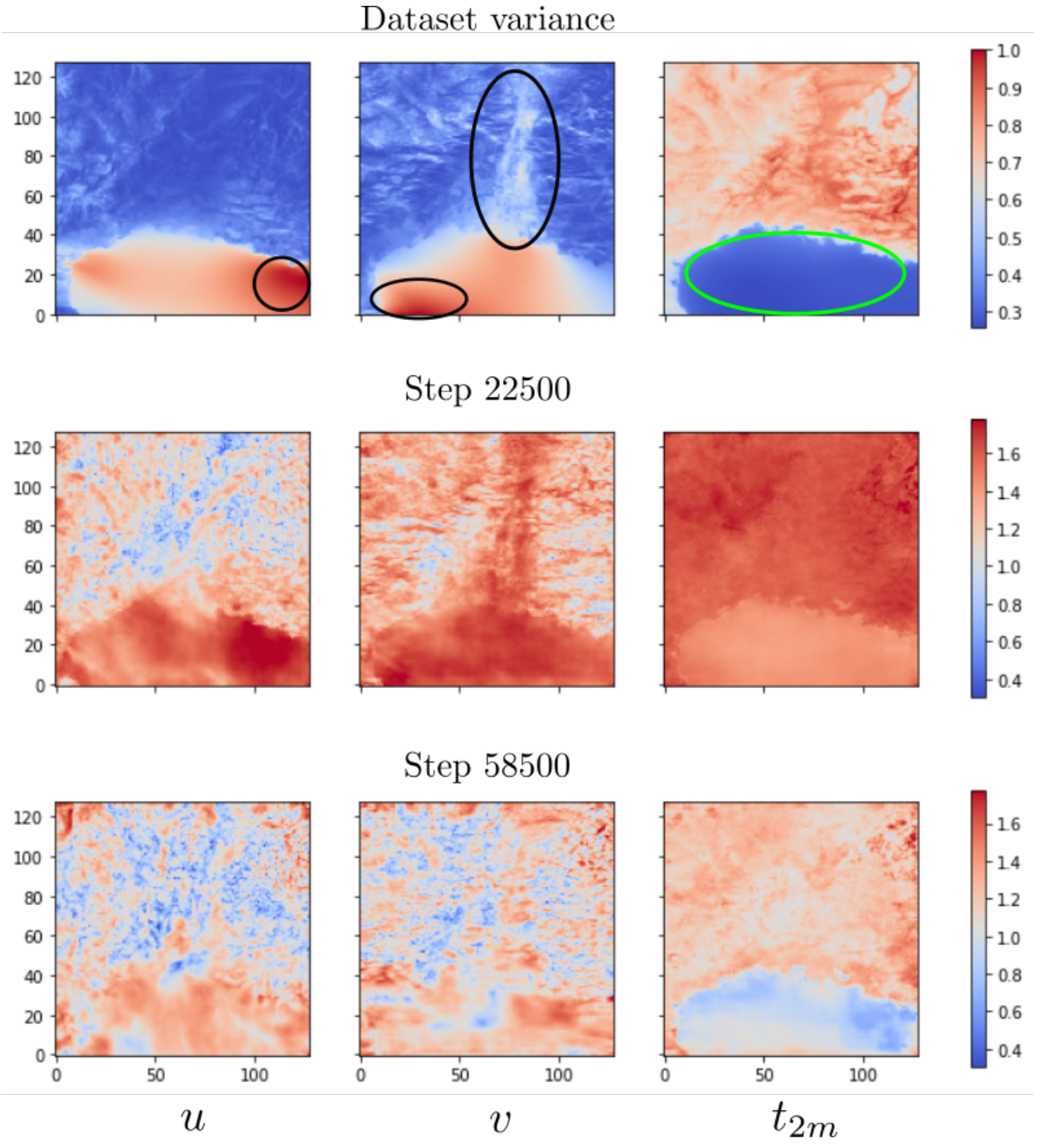

FIG. 8. AROME-EPS per-variable variance maps (top, normalized between 0 and 1) and the pixel-wise EMD maps for two different training steps (middle, bottom). The values of EMD (estimated with 16384 samples) are shown on a common logarithmic scale to emphasize spatial variations and training progression. Regions of higher variance (circled in black in top row) exhibit higher error than others at the beginning of the training (middle row); lower variance regions have lower error (green circling in top row); this difference tends to vanish near training end (bottom row).

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

3) Correlation length scales

The average spectrum of our dataset is almost perfectly fit by our GAN, but it does not take much time for a human observer to distinguish between AROME-EPS and GAN samples. To further analyze the spatial structures of AROME and GAN fields, maps of correlation lengths L_{corr} are shown on Figure 9. These maps show a correct reconstruction of length scales on land, where the location of high and low correlation areas is accurate, and the length scale magnitude order is right. On the contrary, length scales over sea are noisy and exhibit artifacts (checkerboard patterns and border effects), showing a clear gap of quality with respect to land. Note that these artifacts only appear while inspecting this specific metric, while they are either difficult or impossible to spot on individual samples. As will be discussed in Section 6, this is linked to positional information given by specific land patterns. As this information vanishes over sea, non-optimized gradients may show up on this subdomain.

4) Second-order scattering metrics

The s_{21} and s_{22} coefficients are plotted in Figure 10. One can thus compare the distributions of these estimators for GAN samples with the ones of AROME-EPS. At least for small scales, the AROME-EPS dataset is significantly sparser than its GAN counterpart (higher s_{21}). Moreover, AROME-EPS presents sensibly higher s_{22}, indicating it contains more anisotropic, 'filamentary' structures than the GAN samples. The s_{22} 'shape' estimators are rather better fitted by the GAN than the sparsity s_{21} estimators. These observations are consistent with visual inspection of samples. While average spectrograms are nearly indistinguishable for this run, most s_{21} estimators differ significantly. Indeed, the average coefficients are at least one standard deviation away from one another for small λ_1. This difference weakens with larger λ_1, showing that large-scale organisation is better recovered by the GAN.

Both s_{21} and s_{22} distances decrease with training, and this is consistent with the rising quality of GAN outputs (not shown). However, the training dynamics is different from one variable to another. While the sparsity s_{21} distance for t_{2m} is higher than for u and v, the s_{22} distance is lower for t_{2m} than for the wind variables. Globally, this indicates that both s_{21} and s_{22} are reasonable estimators to describe the GANs performance to reproduce the AROME-EPS field's structures.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

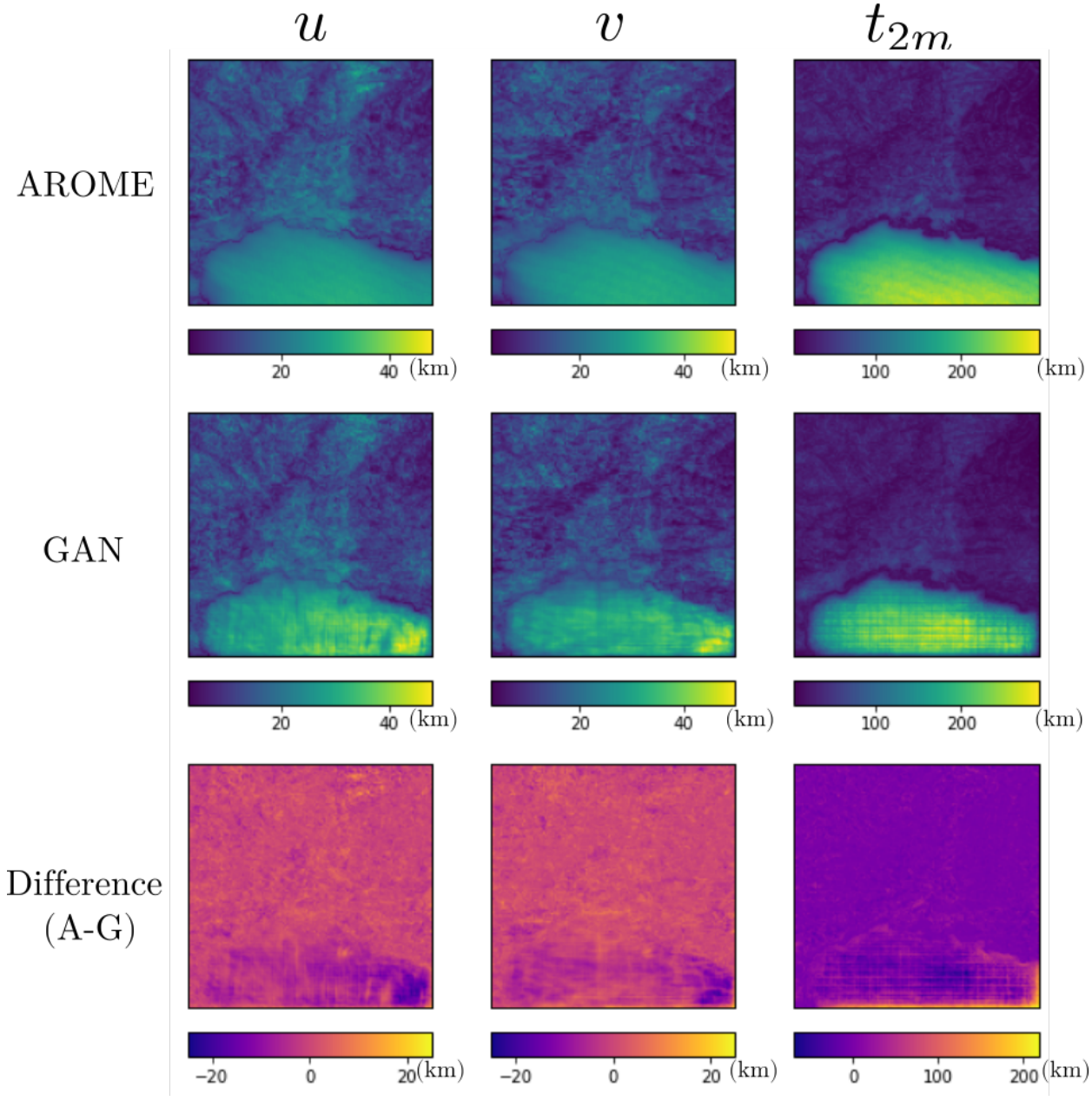

FIG. 9. Correlation length maps for AROME-EPS (top), the GAN (middle) and the AROME-EPS / GAN difference (bottom), for each variable separately. Color scales (in kilometers) are different for each variable but common between AROME-EPS and the GAN.

5) BIVARIATE HISTOGRAMS

Figure 11 shows bivariate histograms of AROME-EPS and GAN samples. A first observation is that the mean and variance of all variables are adequately captured by the GAN. The GAN

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0006.1.

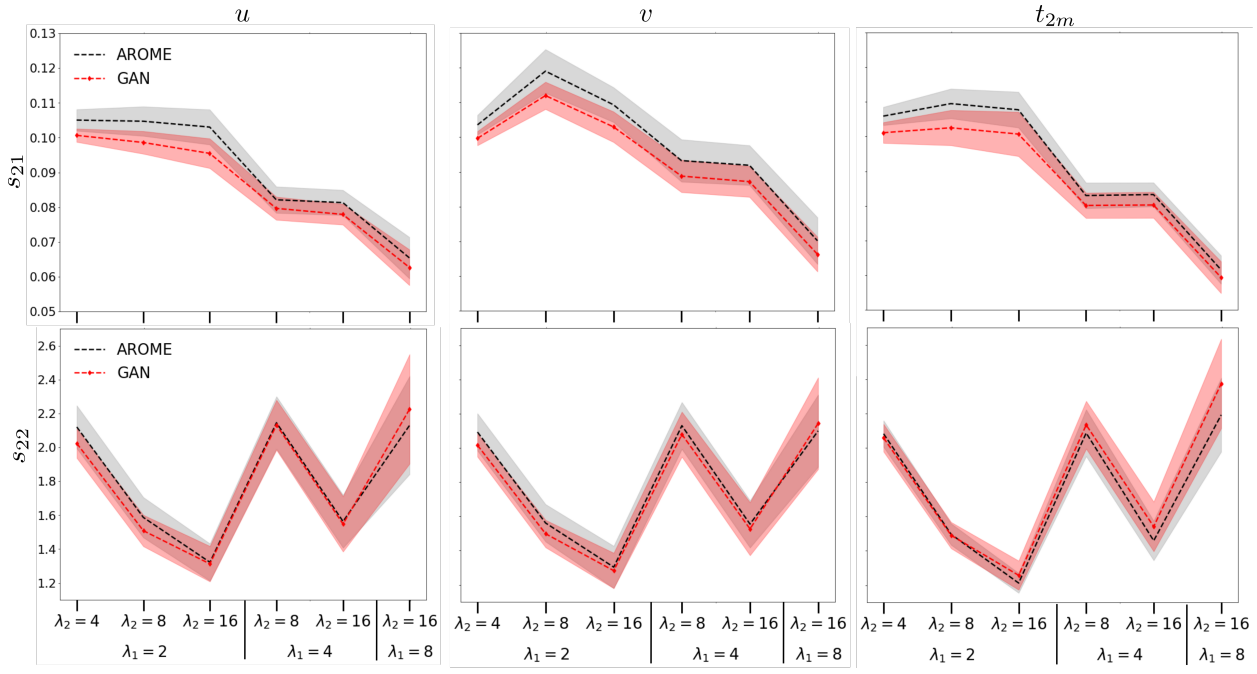

FIG. 10. Scattering s_{21} (top panels) and s_{22} (bottom panels) estimators for u, v, t_{2m}. Scales considered go from $\lambda_1 = 2$ grid points (5.2km) to $\lambda_1 = 8$ grid points (20.8 km) and λ_2 goes up to 16 grid points (41.6 km). Dashed lines represent average quantities, shades represent ± standard deviations.

also surprisingly extrapolates beyond AROME-EPS's data, putting significant probability mass on regions closer to the dataset extremes. Meanwhile, it withdraws mass on other parts of the AROME-EPS distribution. Nevertheless, the logarithmic density scale of the histogram shows that the main modes of the distributions overlap, strengthening the assessment of a correct, global behavior.

6) Percentiles and Inter-Percentiles range

To complete the overview of generation performance, a comparison of percentiles is performed over 66048 samples (so the exact size of the dataset to avoid sampling-related bias). The quantities considered are the 90th and 10th percentiles (Q_{90}, Q_{10}), as well as the 10-90 inter-percentile range (ΔQ). Figure 12 compares the GAN and AROME-EPS statistics. The maximum percentile error of the GAN is limited, but can go up to $3-4\,m\,s^{-1}$ for wind data and 4 K for temperature.

Some regions also show a larger inter-percentiles range for the GAN, mostly over land for wind, and over sea for t_{2m}. Others show a narrower range, mostly the Rhône Valley for u and the

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0006.1.

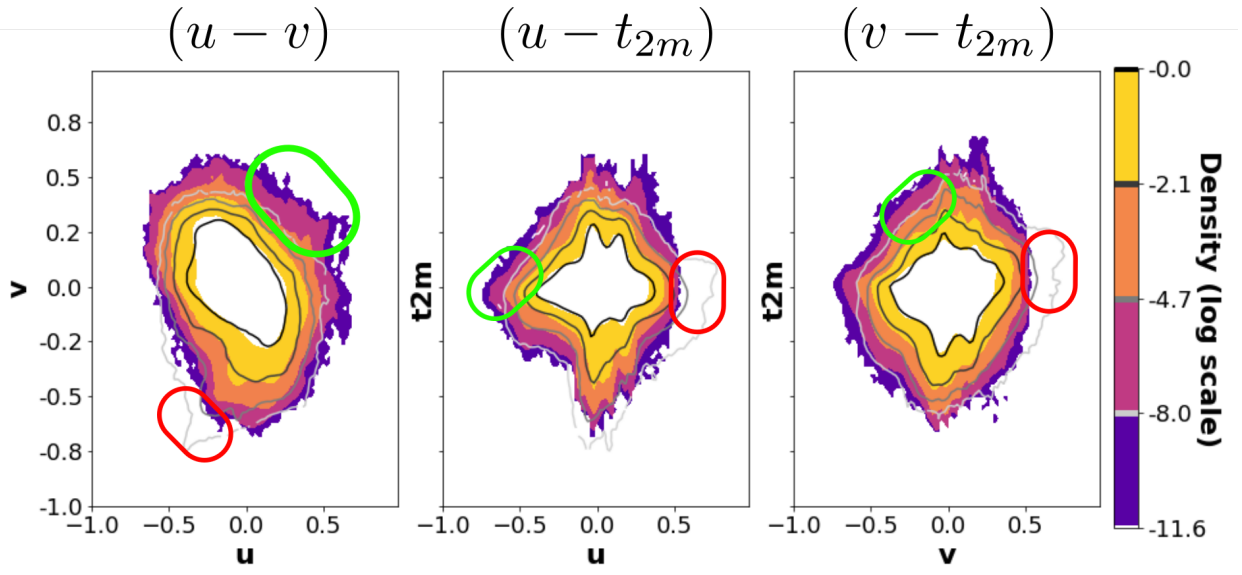

FIG. 11. Bivariate histograms representing the cross-variable correlations. Axes represent the values taken by each variable (on the common, normalized scale used for the training). Contours represent the density of data points for each value, in logarithmic scale: full, colorized contours account for the AROME-EPS distribution ; greyscale contour lines account for the GAN distribution with identical levels to AROME-EPS. Histograms are computed from 16384 samples for each dataset, each pixel of which is one datapoint (2.7×10^5 data points altogether). Note the parts where the GAN extrapolates beyond AROME-EPS (red circling) and the parts where it does not recover AROME-EPS (green circling).

mountains for t_{2m}. The inter-percentile range of the GAN is closer to the one of AROME for temperature than for wind components, where it can be as large as 100%. This supports the fact that the GAN is probably influenced by the positional nature of temperature data.

The average bias of the GAN over all grid points is close to zero for all variables and statistics, except for Q_{10} on t_{2m}. Localized, stronger biases exist however, and they depend on the location as well as on the variable considered. This further supports that the GAN fits the distribution of values, including relatively extreme ones, in an unbiased manner, but can locally exhibit strong deviations from the AROME-EPS distribution, as shown in cross-variable sections.

As a partial conclusion for this section, it has been shown that the GAN has achieved very good quality in terms of sample realism and diversity, power spectrum reproduction, and joint distribution recovery. Moreover, the GAN can generate thousands of samples in a time range of seconds (inference time is around 60s for 16384 samples), making the approach interesting in an

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0006.1.

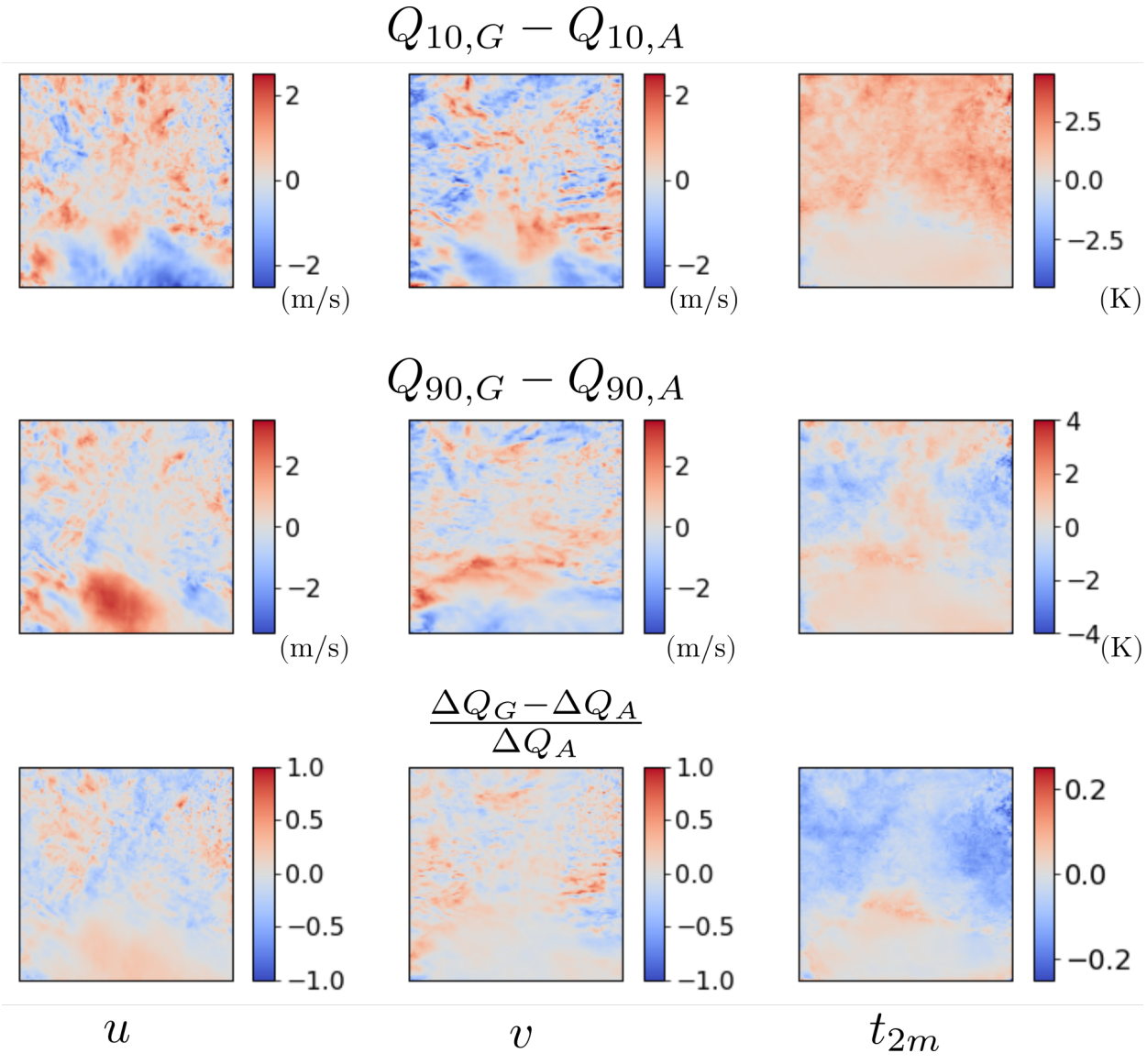

FIG. 12. Difference of percentiles (top, middle) and relative inter-percentile range (bottom) for each variable. Subscript *A* denotes AROME-EPS while subscript *G* denotes GAN. Red nuances positive bias of the GAN with respect to AROME-EPS, blue ones denote negative bias.

ensemble generation framework. Finally, the set of metrics used has been shown to provide a detailed, complementary view of the GAN's capabilities and weaknesses.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

Configuration	Baseline	Config. 1	Config. 2	Config. 3
Generated Variables	(u, v, t_{2m})	(u, v)	t_{2m}	t_{2m}, $orog$
BS/lr_0	$32{,}4 \times 10^{-3}$	$32{,}4 \times 10^{-4}$	$32{,}2 \times 10^{-3}$	$32,\ 2 \times 10^{-3}$

TABLE 4. Summary of the hyperparameters selected for the multivariate experiments.

5. Multivariate configurations: a comparison

The impact of multivariate generation on training has now to be estimated, in order to assess whether adding variables helps the GAN to identify useful correlations, or just makes the task more difficult. The experiment is conducted with 4 different configurations (summed up in table 4):

1. **Baseline** configuration using (u, v, t_{2m}) as generated fields.
2. **Config. 1**: Removing the t_{2m} field and keeping only the generation of the (u, v) couple.
3. **Config. 2**: Removing the (u, v) couple and keeping the generation of t_{2m} field.
4. **Config. 3**: Adding the generation of orography to Config. 2. The constant field of orography is generated by the GAN and also taken into account by the discriminator. This is done to test whether explicitly adding an information related to position adds value for the generation of temperature (which is more correlated to position than the wind).

For each configuration, the best hyperparameters are selected, within the previously used parameter range for lr_0 and BS. This assessment is made on averaging 3 runs' scores with on-the-fly validation metrics ($W_{1,r/c}$ and SWD_{multi}) on 4096-sample batches, completed by visual inspection of samples. This allows for a fast and reliable selection of hyperparameters, which are summed up in Table 4. Once these are selected, evaluation is performed on another set of 3 runs for each selected configuration. This final evaluation is performed with batches of 16384 samples, and all the previously described metrics are used to yield the most extensive evaluation. The results are reported in Tables 5 and 6.

Tables 5 and 6 show that a general effect of reducing the number of generated variables is an increasing performance on most metrics related to spatial consistency (PSD, correlation lengths, scattering metrics). Another interesting pattern is that scores of Config. 3 (t_{2m} and orography) are generally worse than those of Config. 2 (t_{2m} only), and sometimes even than Baseline. Adding orography information thus seems to have a mixed effect. On the one hand, it degrades the

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

Metric	$W_{1,r}$	$W_{1,c}$	PSD_u	PSD_v	$\text{PSD}_{t_{2m}}$	SWD_{128}	SWD_{64}	SWD_{32}	SWD_{16}	$s_{21,u}$	$s_{21,v}$	$s_{21,t_{2m}}$	$s_{22,u}$	$s_{22,v}$	$s_{22,t_{2m}}$
Unit/Scale	$\times 10^{-3}$		$\times 10^{-1}$ dB			$\times 10^{-3}$				$\times 10^{-3}$			$\times 10^{-2}$		
Baseline	13	12	8.1	8.9	11	5.7	7.3	12	39	5.0	3.8	5.6	4.7	4.6	1.7
Config. 1	*14*	12	**7.4**	**7.9**	NA	*21*	*20*	*22*	*59*	**1.6**	**0.8**	NA	**2.7**	**2.7**	NA
Config. 2	**11**	**11**	NA	NA	**7.4**	**6.6**	*10*	**10**	**30**	NA	NA	**0.5**	NA	NA	**0.7**
Config. 3	**11**	**10**	NA	NA	*13*	*7.8*	*10*	12	**19**	NA	NA	**3.2**	NA	NA	**1.2**

TABLE 5. Global score card to compare multivariate experiments. Reported scores correspond to the average best score obtained after training saturation for the 3 runs. For all metrics considered, lower is better. Better scores with respect to baseline are shown in bold black, while worse scores are in italic. NA: "not attributed", is used when the metric is not applicable to the configuration.

	$\text{MAE}(L_{corr,u})$	$\text{MAE}(L_{corr,v})$	$\text{MAE}(L_{corr,t_{2m}})$
Baseline	2.4 km	2.2 km	11.7 km
Config. 1	**1.8 km**	**1.6 km**	NA
Config. 2	NA	NA	*13.6 km*
Config. 3	NA	NA	*15.2 km*

TABLE 6. Mean absolute error for correlation length maps. Reported scores correspond to the best average score obtained after training saturation for the 3 runs. For all metrics considered, lower is better. Better scores with respect to baseline are shown in bold black, while worst scores are in italic. NA: "not attributed" is used when the metric is not applicable to the configuration.

synthesis of temperature spatial structures, as emphasized by PSD error, correlation length error and scattering metrics. On the other hand, $W_{1,r/c}$ scores, as well as the SWD_{16} scores, are sensibly improved when orography is added. Removing temperature and orography and keeping wind variables has an opposite effect. Indeed, Config. 1 shows improved spectral, scattering and correlation lengths metrics, while $W_{1,r/c}$ scores slightly degrade and SWD_{multi} scores dramatically degrade. This shows a lack of ability of the GAN to capture the diversity of patterns in the dataset, while improving the quality of individual samples. The largest distributional discrepancy of Config. 1 even hints at some form of mode collapse.

Altogether, the quality of samples also improves when comparing Config. 1 and 2 to Baseline (Figure 13). This is consistent with the majority of metrics involving spatial consistency. Especially scattering metrics show a drastic improvement when reducing the number of variables. The GAN is therefore much more able to identify and generate multi-scale organisation in the samples, albeit

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

at the expense of pattern diversity. This points at the GAN using cross-variable correlations to improve the diversity of samples, rather than their quality.

6. Learning absolute grid-point position: analysis and consequences

The above experimental results give a set of observations that can be exploited to diagnose the strengths and flaws of the training design and their interaction with neural network architecture.

Let us summarize some of them:

1. The error signal at the beginning of the training is strongly correlated to the pixel-wise variance of the dataset (Section 4).

2. Large-scale EMD are far worse than small-scale EMD in all configurations (Sections 4 and 5).

3. Performance for correlation length scales is far better on land than over the sea (Section 4).

Given that learning is performed on a fixed spatial domain, it is very likely that the main source of information for learning in our setup is the implicit encoding of absolute grid-point position. This phenomenon is already acknowledged in literature for convolutional networks (Alsallakh et al. 2021; Zhang 2019), and has been extensively studied in the case of GANs by Xu et al. (2020). It is usually explained by the use of padding in convolutional layers: adding rows and columns of zeros in the intermediate layers allows the network to detect the boundaries of the feature maps, and thus implicitly infer the position of each pixel.

The present setup goes a step further by adding variables that are more or less directly correlated to surface state, and thus to absolute grid-point position. Temperature's variability is mostly position-related over land, as it is obviously the case for orography and also, although moderately, for 10-m wind. Over sea, this position-related information fades out, while transient features, such as wind fronts, are prominent: the bias is weaker and the GAN struggles generating correct correlation structures.

This positional bias probably plays a key role of dedicating most of the networks' power to extract and fit position-related features. Hence, this analysis is a plausible explanation for another set of observations:

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

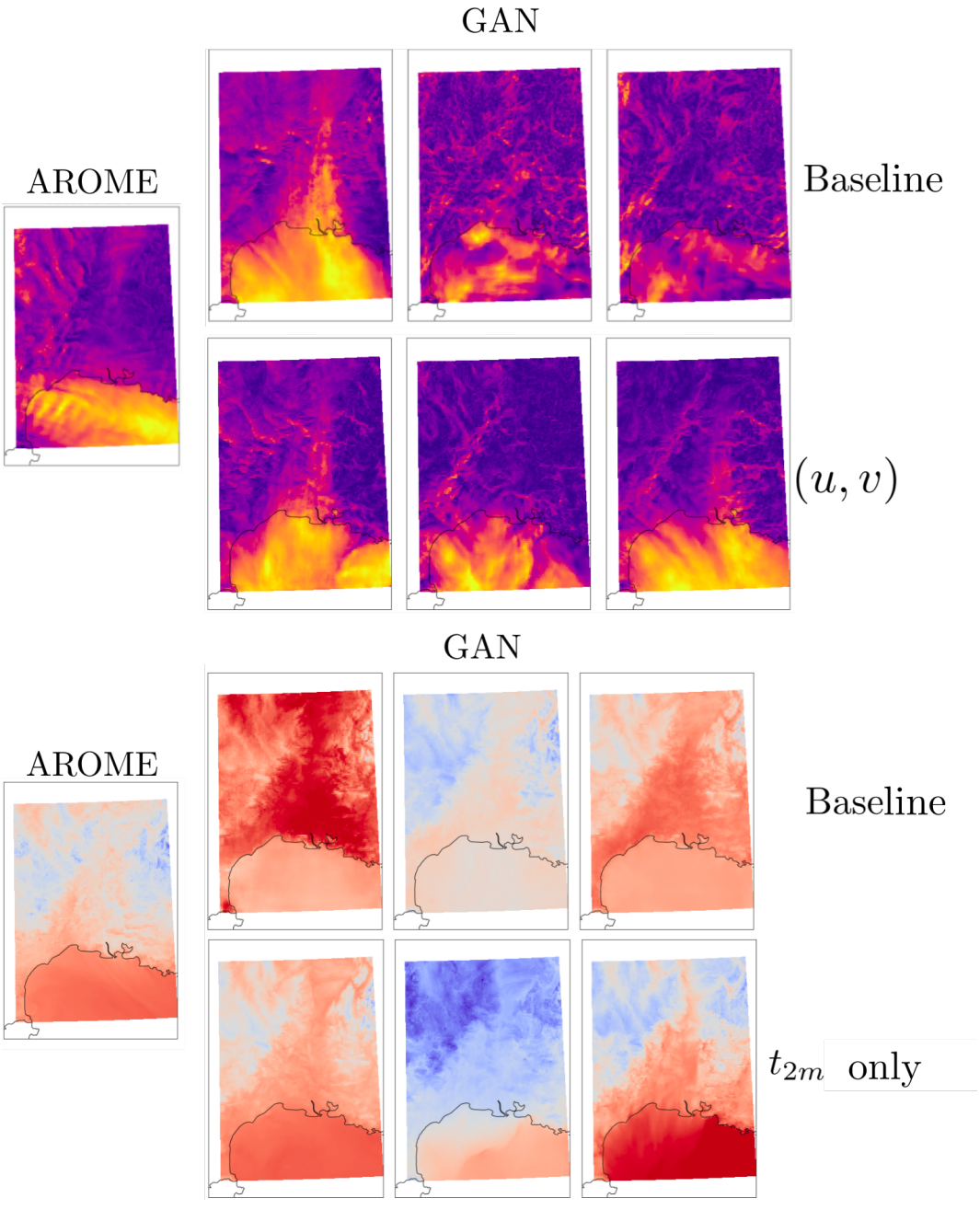

FIG. 13. Comparison of random samples from the GAN when shifting the training configuration. One random AROME-EPS sample is left for comparison. The most successful configurations for Config. 1 (wind variables only) and Config. 2 (temperature only) are shown. In both configurations, most quality-related metrics do increase. The organisation of long-range structures is enhanced in both Config. 1 and Config. 2, with fronts more visible and showing better long-range structuring. Value scale here is voluntarily left free, to enable visual-only comparison.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

1. Right from the Baseline configuration (where no explicit position is given to the network), the reconstruction of temperature correlation with altitude is very accurate. In this case, learning position-related features with fine-grained spatial detail is largely helpful.

2. Adding orography as a constant field to generate is largely detrimental to the sample quality, according to the scores used, but improves the largest scales of multi-scale SWD (Section 5). Positional information at large scales gives information on the overall structure of the field, and is then most useful to generate the right distribution of patterns. Conversely, reinforcing this bias through orography accelerates position-based overfitting.

3. The GAN is not able to use cross-variable correlations to improve individual sample quality, but maintains more diversity (Section 5): the added information of each variable in the Baseline configuration is likely redundant if it is only position-related. This prevents the GAN to improve by using the less redundant, transient features which differentiate the 3 variables.

4. The GAN trades off diversity for quality in the wind-only configuration (Section 5). This configuration is the one where positional information has least weight. It is thus probable that reducing the positional bias makes the GAN focus on transient features quality that play a larger role in discrimination, while relaxing the diversity constraint and narrowing the distribution of patterns. This is likely guided by large-scale pattern detection being harder without positional bias, as evocated in point 2.

These explanations imply that the setup is prone to overfitting, and that, contrary to the common assumption, increasing batch size will degrade the performance of the GAN. We thus conduct a final analysis using different initial learning rates and batch sizes ($BS \in \{32, 64, 128, 256, 512\}$, $lr_0 \in \{4\times10^{-4}, 2\times10^{-3}, 4\times10^{-3}\}$). Using the Baseline configuration, we train the GAN from scratch for each pair of learning rate and batch size, up to loss saturation. We first observe that loss saturation occurs earlier with increasing batch size (cf Appendix C), underpinning the above hypothesis. Figure 14 shows the relative degradation/improvement of metrics with respect to the $BS = 32$ configuration for each learning rate. Once saturated, $W_{1,c}$ only slightly increases with batch size, at all learning rates. On the other hand, quality metrics such as PSD and s_{21}/s_{22} errors drastically degrade when batch size increases (up to $\times5/\times10$ degradation for PSD). SWD_{avg}

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

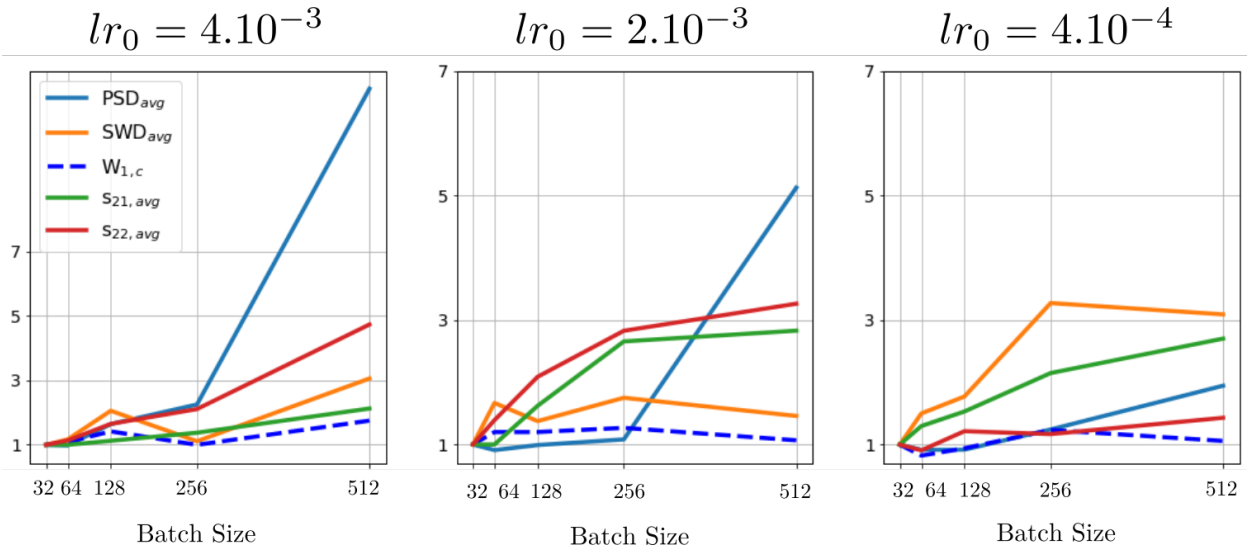

FIG. 14. Relative evolution of different scores with respect to their value at the $BS = 32$ configuration, for different, decreasing learning rates.

follows a path in between. The effect is less pronounced with diminishing learning rates, but it was observed that smaller learning rates degrade scores globally (cf. Appendix C).

Reproducing pixel-wise, variable-wise distributions is thus rather an "easy" mode of convergence, achievable for most GAN configurations. Increasing batch size then likely strengthens the position-learning dynamics. The GAN rapidly memorizes position-related features and more or less forgets about the transient structures, which are smoothed out by large batches.

This set of explanations is consistent with the hypothesis of absolute grid-point position learning. It also echoes 'classical' quality-diversity trade-offs encountered in GANs (Radford et al. 2015; Zhang et al. 2019; Brock et al. 2018) and the general perception-distortion trade-off of generative models (Blau and Michaeli 2018) ; in our case, the trade-off is balanced by the variables fed to the GAN, and the amount of positional information they contain. Learning on a fixed domain is a crucial component of the present setup, which explains both the good quality of individual samples and grid-point distributions, as well as the scale-dependent diversity fit.

Whether such positional bias *should* be learnt is an open question in general. As was shown by temperature correlation with altitude, this is very helpful in weather-related tasks, where many features depend on the absolute position. The way it is added in the training process, through architectural features and the data itself, can thus be clearly framed and controlled with *a priori*

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0006.1.

heuristics. As an example, the positional bias could be strongly attenuated with variables that are much less dependent on absolute position (e.g, temperature at 850 hPa). This could also be the case if the networks were trained on random domain patches while conditioning both networks through orography. In this case, the task is strictly more difficult as the diversity of the dataset increases. We performed some preliminary experiments in this randomized setup: while it seems that one indeed gets rid of positional bias (notably, increasing batch size improves SWD_{multi} and PSD errors), this remains to be detailed and confirmed in future work.

It is also possible that more sophisticated architectures such as ProGAN (Karras et al. 2018) and StyleGAN (Karras et al. 2019, 2020), which explicitly handle scale-dependent pattern generation and disentangle features, perform notably better on the same setup.

7. Conclusion and perspectives

In this paper, meaningful metrics have been developed and applied to assess the ability of a GAN to emulate outputs of the kilometer-scale AROME-EPS weather forecasting system. From the above evaluation, the following conclusions can be drawn:

1. Multiple metrics, borrowed either from weather science or computer vision, are necessary to diagnose the ability of a GAN to consistently generate weather states. Namely, going beyond a mere spectral analysis to describe the quality of generated samples proved useful. Multi-scale SWD was successful in characterizing the diversity evolution with scale ; scattering coefficient were used to assess the consistency of structures, while the local correlation length scales enabled a position-wise analysis of correlation reconstruction.

2. A residual GAN architecture is plainly able to generate multivariate distributions of NWP models at kilometer scale. In particular, it can reproduce detailed textures as well as long-range events with a good diversity. State-of-the-art regularization techniques and networks are necessary for this task, and their training parameters must be carefully set to avoid divergence.

3. A study on multivariate generation was performed to probe the effects of adding and removing variables to the training setup. An important phenomenon happening in our GAN was characterized: the positional bias, induced both by padding and surface variables such as 2-meter temperature, is a prominent drive of the learning process. This is a double-edged

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

sword: it enables fast convergence and emulation of crucial features such as temperature correlation with altitude, while it degrades the ability of the GAN to generate high-quality transient structures and accelerates the occurrence of overfitting. This seems to be part of the wider quality-diversity trade-off, whose major component in the present case is given by the specific weather variables used.

Precipitation has not been addressed in this work, because of the extremely skewed nature of its distribution, specifically the overwhelming class of zero-precipitation days. Resampling techniques (Sha et al. 2020; Ravuri et al. 2021) would arguably tackle this point, but the relatively small size of the database discouraged us to go further in this direction on a first trial. This is a natural path for future work.

Another promising path is the generation of states based on the current weather situation. This framework has been used by many downscaling studies (Leinonen et al. 2021; Harris et al. 2022) that take low-resolution data as inputs to generate ensembles of high-resolution outputs. One could then ask whether a GAN framework could be used to increase the size of operational ensemble forecasts, at a minimal computing cost. Open challenges would then be the precise way to condition the GAN with ensemble outputs at the same resolution, as well as the ability to control GAN-produced outliers. We believe the results showed in this study are encouraging enough to go further in this direction.

Finally, a largely unexplored path is the production of temporal sequences of forecasts at the lead times usually covered by the operational NWP models (up to 48-72h). While it remains open whether the GAN framework is adapted to such a task, this would be a necessary step in order to use data-driven, high-resolution, real-time ensemble emulation.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

Acknowledgments. We deeply thank Léa Berthomier and Bruno Pradel for their thorough technical support with Météo-France's computing infrastructure. We thank Camille Besombes, Olivier Pannekoucke, Ronan Fablet, Arnaud Mounier and Thomas Rieutord for insightful discussions. This work was performed as part of the ANR project 21-CE46-007 "Probabilistic prediction Of Extreme weather events based on ai/physics SYnergy (POESY)" led by one of the authors (LR). One of the authors (CB) performed this work during his PhD, funded by the French Ministère de la Transition Ecologique as part of the PhD program of the "Ingénieurs des Ponts, Eaux et Forêts" civil Corps.

Data availability statement. The code used to train networks and analyze data through all metrics can be found at https://github.com/flyIchtus/multivariate-GAN. The AROME-EPS dataset used in this study will be made available at the end of the ANR project.

APPENDIX A

Implementation details

The implementation of our GAN makes use of several techniques, acknowledged to either facilitate convergence or accelerate computing. Here are reported the ones that were helpful. The code is written with PyTorch (Paszke et al. 2019), using the multi-GPU Horovod API (Sergeev and Balso 2018).

1. The residual blocks we use follow usual guidelines of literature (Miyato et al. (2018), Besombes et al. (2021), Ravuri et al. (2021)). The main block consists of two stacked 3×3 convolutions followed with LeakyReLU and BatchNorm, with a bilinear upscale/downscale layer. Either a 1×1 convolution or a direct sum is used as residual shortcut.

2. We use Automatic Mixed Precision (AMP), casting most operations to half-precision. This leads to a dramatic acceleration of training and slashes memory consumption, keeping all runs below 12 hours and leaving space for later development of architectures. Unfortunately, this also comes with stability issues: some of the runs produced NaNs at their very beginning, with specific hyperparameter configurations being completely hampered by AMP while running smoothly with simple precision.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

3. It was found that the discriminator's gradients or the failed cases were violently oscillating at the beginning of training. We then introduced a small warm-up procedure where the discriminator was updated several times for one update of the Generator. Choosing a update ratio of 5 as in Gulrajani et al. (2017), on the single first generator step considerably reduced the oscillations of gradients and made training stable for most of the 180 runs conducted for this study.

4. Initialization of the neural networks weights has been shown an important factor for training convergence (Bengio and Glorot (2010)). Here we keep the default random initialization for all linear layers, while using Orthogonal Initialization for all convolutional layers. Besides generally having a beneficial impact on training (Saxe et al. (2014)), this naturally helps the Spectral Normalization regularization we adopted by starting with already spectrally-normalized (random) weights.

APPENDIX B

Detailed description of metrics

a. Pixel-wise earth mover distance

For a distribution with only one degree of freedom, Wasserstein (earth mover) distance amounts to comparing cumulative distribution functions through the integral:

$$W_1 = \int |F_{\mathbb{P}}(x) - F_{\mathbb{Q}}(x)| dx$$

Being given two series of N *sorted* samples S_p and S_q drawn from $\mathbb{P}$ and $\mathbb{Q}$ respectively, computing this integral comes down to averaging the difference of sorted values:

$$W_1 \approx \frac{1}{N} \sum_i |S_{p,i} - S_{q,i}|$$

The computing complexity is essentially due to sort ($O(N \log N)$). The absolute value of W_1 depends on the unit of the data, hence on the normalization process. To give a view of what absolute values mean in our case, figure B1 presents two pixels with different W_1 distances and

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

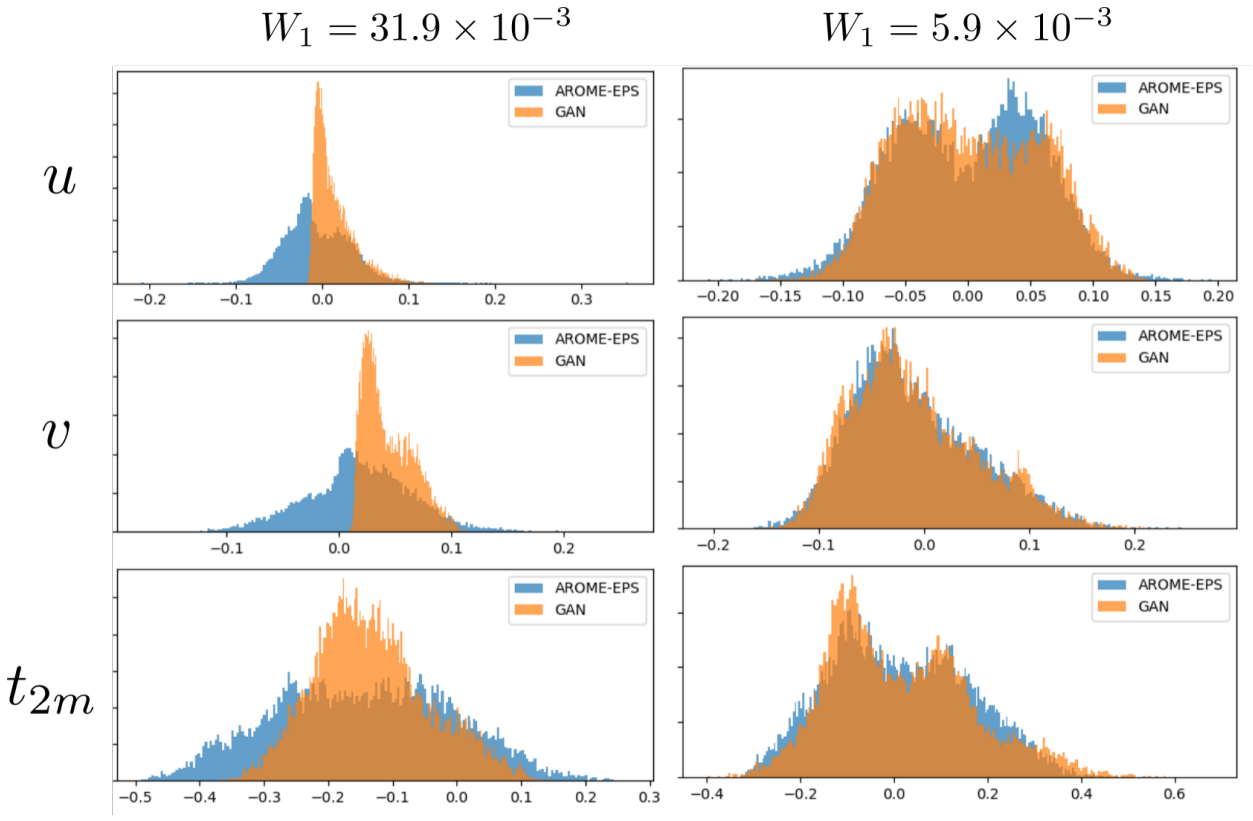

FIG. B1. Correspondence between Wasserstein distance estimation and distributional fit for two grid points in the best-performing configuration. Left: grid point with poor fit, right: correctly fitted grid-point. Each distribution is estimated from 16384 samples.

gives a correspondence to the shape of GAN and AROME-EPS distributions. Poor distributional fit (misplaced density, poor reconstruction of bimodal data) is characterized by a rather high W_1 value, while a good fit (correct spread and tails, bimodality captured) shows lower W_1. This can be compared to the averaged $W_{1,r/c}$ obtained by the best-performing configuration (around $12-13\times10^{-3}$).

b. Sliced Wasserstein distance

Here we draw extensively from two works from Rabin et al. (2011) and Karras et al. (2018). The Wasserstein distance is known to be an informative metrics for distributions (Arjovsky et al. 2017), yet is computationally intractable and exhibits large variance in high-dimensional spaces (Ramdas et al. 2015). Monte-Carlo approximation of such a distance is defined in Rabin et al. (2011), and is

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0006.1.

an unbiased and robust way to cope with the burden of W_1 estimation in high-dimensional spaces (Kolouri et al. 2018).

For multi-level estimation, we first decompose the original field into a laplacian pyramid, from finest to coarsest scales. The process then consists in selecting a batch of random neighbourhoods of 7×7 pixels, and normalizing each variable of the samples with respect to batch and spatial mean and standard deviation. The distributions of neighbourhoods from the GAN and from AROME-EPS samples are finally compared with the help of SWD. Since neighbourhoods include several pixels, they have several degrees of freedom: SWD is thus a way to estimate the multidimensional EMD on these neighbourhoods. We use the parameters of Karras et al. (2018) without modification, with 512 unit directions for SWD and 128 random neighbourhoods for each level.

c. Scattering coefficients

Obtaining scattering coefficients consists in successive convolutions with wavelet filter banks $\{\psi_\lambda\}_{\lambda \in \{2^0, \ldots, 2^J\}}$. The λ index corresponds to the discrete scale of the filter bank, i.e the number of pixels in the filter's bandwidth. The largest scale probed, denoted by index J, typically corresponds to a half of the spatial extent of the field. In order to treat 2-dimensional fields, angular dependency is added to the family of wavelets ψ, so λ indexes both scale and direction θ within the $[0, \frac{\pi}{8}, \ldots \frac{7\pi}{8}]$ discrete interval: $\lambda = (2^j, \theta)$. We consider the common family of complex Morlet wavelets, satisfying stability and invertibility constraints (Mallat 2012), and we use the Python package *Kymatio* developed by Andreux et al. (2018).

The convolution of a field X with this wavelet family yields a set of *first-order* scattering maps $M_1(\lambda_1)$:

$$M_1(\lambda_1) = |X \star \psi_{\lambda_1}|$$

The convolution on a given λ_1 identifies structures of typical length scales λ_1. The modulus then provides robustness of the transform to local deformations: slightly deformed patterns at the scale of λ_1 produce close values of $|x \star \psi_{\lambda_1}|$. Combining convolution and modulus yields first-order scattering images. These images themselves exhibit specific structures that vary on several (larger) scales: *second-order* maps can also be extracted at scale $\lambda_2 > \lambda_1$:

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

$$M_2(\lambda_1, \lambda_2) = ||X \star \psi_{\lambda_1}| \star \psi_{\lambda_2}|$$

These second-order maps represent the organisation of λ_1 structures at the scale λ_2. Maps can then be spatially averaged to produce global, translation invariant coefficients. Namely:

$$S_1(\lambda_1) = \langle M_1(\lambda_1) \rangle_{space}$$
$$S_2(\lambda_1, \lambda_2) = \langle M_2(\lambda_1, \lambda_2) \rangle_{space}$$

(where $\langle \cdot \rangle_{space}$ denotes spatial-averaging). As emphasized by Cheng et al. (2020), S_1 coefficients are similar to spectral density as they decompose the signal scale by scale, and then average over spatial dimension. Second-order scattering coefficients probe the organisation of each scale.

As a reminder, the summary statistics we use are drawn from Cheng and Ménard (2021). They compare the amount of information stored in different scattering coefficients. Signal 'sparsity' is probed through an orientation-averaged comparison between second and first-order coefficients:

$$s_{21}(\lambda_1, \lambda_2) = \left\langle \frac{S_2(\lambda_1, \lambda_2)}{S_1(\lambda_1)} \right\rangle_{\theta_1, \theta_2}$$

While distinction between roundish and filamentary shapes (accounting for anisotropy) is better probed with a ratio of colinear versus orthogonal orientations for second-order coefficients:

$$s_{22}(\lambda_1, \lambda_2) = \left\langle \frac{S_2(\lambda_1, \lambda_2)_{\theta_1 = \theta_2}}{S_2(\lambda_1, \lambda_2)_{\theta_1 \perp \theta_2}} \right\rangle_{\theta_1}$$

To illustrate our claims, we take a subset of 256 AROME samples of wind speed, and we generate Gaussian noise fields from the exact same spectrum. The created Gaussian maps have little multi-scale organisation, and AROME samples are thus "sparser" than their Gaussian counterparts. Figure B2 shows that while spectra are well-aligned, there is a significant difference in field's structure, as shown by the s_{21} coefficient discrepancy. AROME is also slightly less isotropic, as shown by its higher s_{22} for the smallest λ_1. This correlates well with the visual examination of samples.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

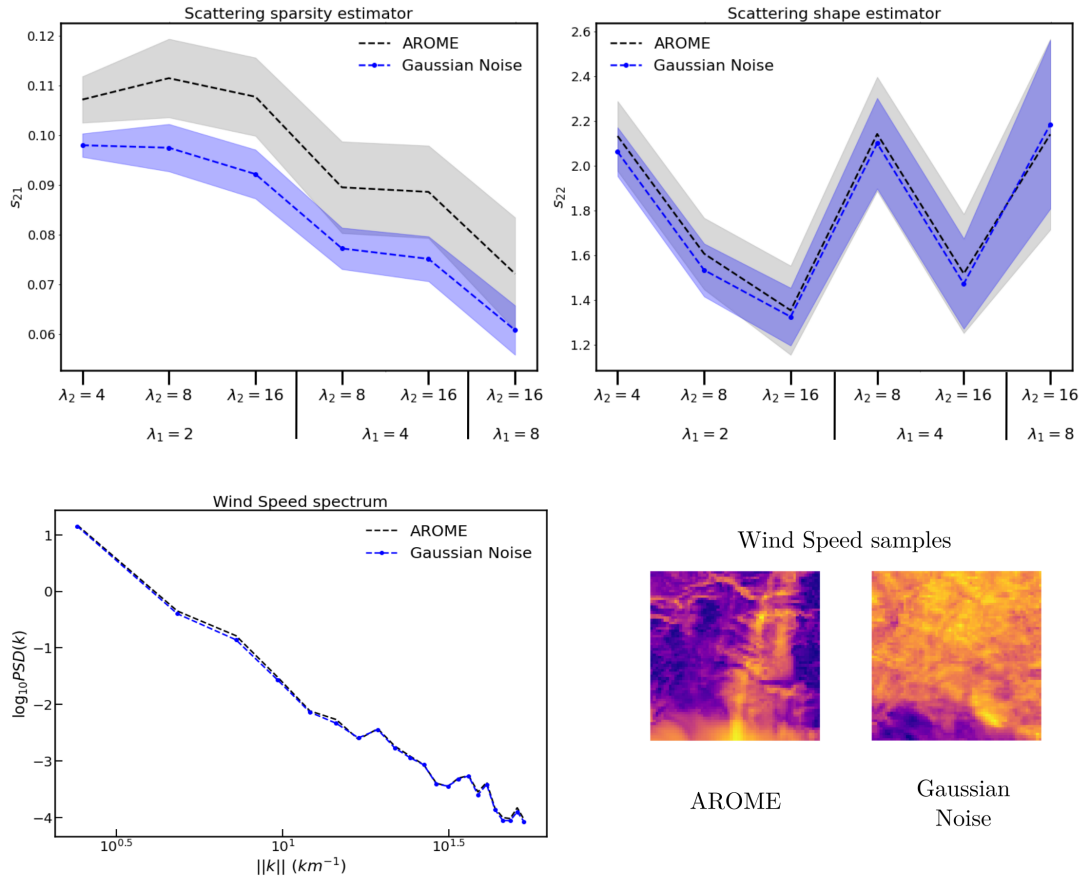

FIG. B2. Comparison of Gaussian noise and AROME wind speed with respect to scattering estimators s_{21} and s_{22} (top), and to power spectral density (bottom). Samples are shown to provide visual assessment. Dashed lines correspond to average quantities, while shades correspond to ± standard deviation. The scales shown go from 2 to 16 grid points.

To provide estimates of absolute values for RMSE on scattering coefficients and SWD_{multi}, Table B1 summarizes the scores obtained by the Gaussian field maps and the Baseline configuration. Note that since the samples are normalized before applying the RMSE, all variables share a common scale of values. This table shows that the GAN performs sensibly better than a Gaussian field in any configuration, and that a Gaussian field does not recover the correct distribution of patterns as measured by SWD_{multi}.

APPENDIX C

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

Metric	SWD_{128}	SWD_{64}	SWD_{32}	SWD_{16}	$s_{21,avg}$	$s_{22,avg}$
Unit/Scale	$\times 10^{-3}$				$\times 10^{-3}$	$\times 10^{-2}$
Lower bound (AROME vs AROME)	1.5	1.5	1.6	4.6	0.0	0.0
Gaussian field	53	60	49	316	9.0	9.7
Baseline	5.7	7.3	12	39	4.8	3.7

TABLE B1. Comparison between scores against AROME-EPS dataset for Gaussian field and the Baseline GAN configuration. Scattering estimators have been averaged over all three variables for Baseline configuration.

Hyperparameter search

Experiments were carried out on 5 different batch sizes ($BS \in \{32, 64, 128, 256, 512\}$) and 3 different initial learning rates ($lr_0 \in \{4\times10^{-4}, 2\times10^{-3}, 4\times10^{-3}\}$). Each configuration was run 3 times to account for training variablity.

For all configurations, the discriminator loss curves exhibit a deep trough followed by a slower ascent, up to a value below 2.0, after which the loss plateaus ; meanwhile, the generator loss produces bumps after an abrupt decrease, before oscillating around 0.0 when the D loss reaches its maximum level. Figure C1 exposes some examples of this behaviour. The converged regime corresponds to situations where D is unable to separate the AROME-EPS and the GAN samples: it is likely to indicate convergence of the algorithm on a local minimum. Using $BS \geq 256$ provokes rapid saturation of the losses, indicating early stagnation of learning. Reducing the learning rate attenuates this effect and lengthens the ascent part. The $BS = 32$ experiment did not reach the plateau for any of the learning rates but the highest, indicating that batch size and learning rate both control the learning speed. Another direct effect of batch size increase is the reduction of loss oscillations.

At the point where the D loss reaches saturation, training is completed for most runs, as our control metrics often saturate (not shown). In some cases, some components of SWD_{multi} tend to slighlty increase after the plateau, indicating possible overfitting. A score card for all the tested hyperparameter configurations can then be drawn. For each configuration, the best (i.e lowest) score obtained is selected, for all metrics, after D loss saturation when it happens, or the best score altogether. Results are reported in Table C1.

Trying to select the best-performing configuration from this table, one can rule out the $BS = 512$ configurations, mainly because of high PSD errors. This corresponds to a very low visual quality of samples (blurry images with checkerboard artifacts). Small batch sizes seem to perform better

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

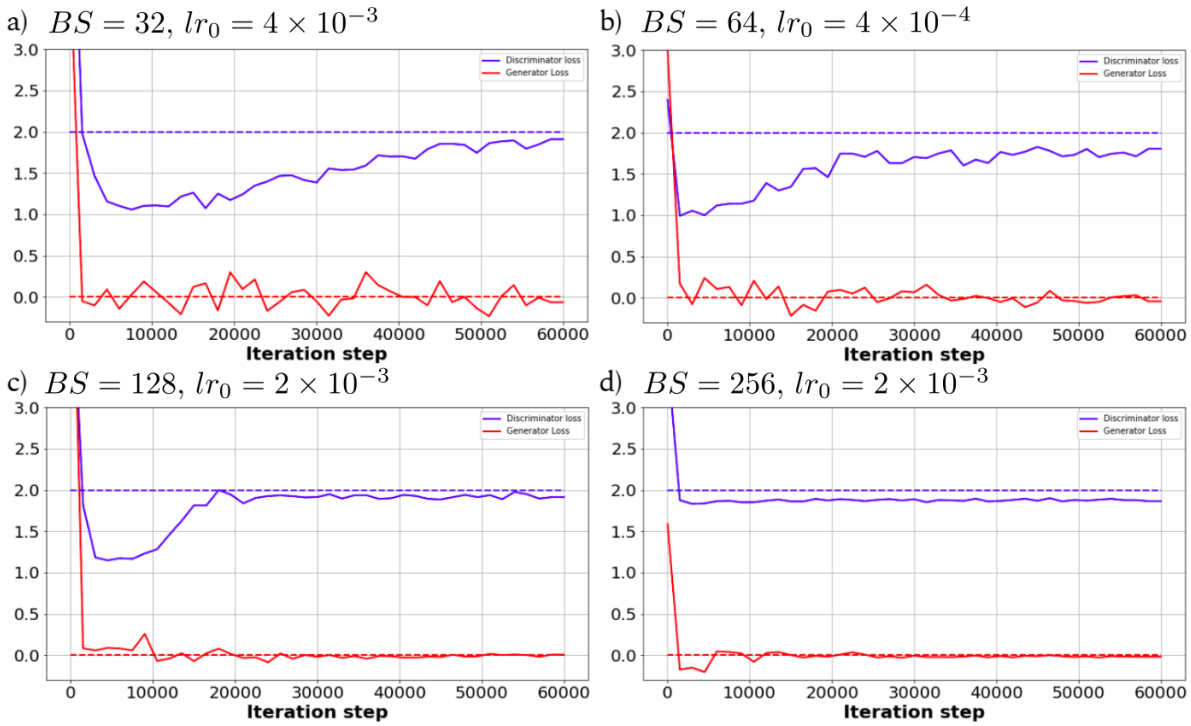

FIG. C1. Typical patterns observed in training. The network losses (blue: discriminator, red: generator) are represented along with the number of steps, for different batch sizes and initial learning rates. Discriminator losses tend to converge near 2.0 (blue lines), while generator losses oscillate around 0.0 (red lines).

than others, especially with respect to EMD and SWD metrics, even if they do not always perform best on PSD errors. Moreover, the score spread for each configuration is rather wide and it is not rare that different configurations produce scores on overlapping ranges. Altogether, $lr_0 = 4 \times 10^{-3}$ seems to perform better than any other, with the $BS = 32$ configuration scoring best on a wide range of metrics.

Only the most successful configurations of Table C1 are kept to get a view of their EMD-scattering scores and check their ranking. These observations are summed up in Table C2. The ranking slightly changes: while increasing batch size degrades the scores similarly to previous experiments, intermediate learning rate of 2×10^{-3} produce the best scores obtained. All experiments that are not reported in Table C2 show worse scores than the ones shown.

APPENDIX D

Does the GAN copy the dataset?

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0006.1.

BS, lr_0	$W_{1,r}$	$W_{1,c}$	PSD$_u$	PSD$_v$	PSD$_{t_{2m}}$	SWD$_{128}$	SWD$_{64}$	SWD$_{32}$	SWD$_{16}$	SWD$_{avg}$
Unit/Scale	$\times 10^3$		$\times 10^{-1}$ dB			$\times 10^3$				$\times 10^3$
$32, 4\times10^{-3}$	**13/12/13**	**12/12/12**	**8.1/7.4/9.7**	8.9/8.1/9.8	11/10/11	**5.7/5.1/6.2**	**7.3/6.5/7.7**	12/10/15	**39/34/48**	**18/17/21**
$64, 4\times10^{-3}$	14/13/16	13/12/15	8.2/7.9/8.3	**8.3/8.2/8.5**	11/9.4/12	**5.8/5.1/6.1**	8.0/6.5/9.7	12/10/16	51/38/70	21/16/27
$128, 4\times10^{-3}$	19/16/21	17/15/19	11/8.7/14	16/9.3/28	19/12/31	13/8.9/19	20/12/30	27/18/40	77/71/87	37/30/43
$256, 4\times10^{-3}$	14/11/19	12/9.3/13.8	17/20/23	23/12/28	23/10/30	11/7.5/14	**7.2/5.7/9.9**	**10/6.5/17**	50/31/78	20/15/43
$512, 4\times10^{-3}$	*23/21/28*	*21/18/26*	*91/56/110*	*113/78/134*	*133/125/148*	*65/57/79*	39/32/51	35/17/68	78/60/102	55/46/71
$32, 2\times10^{-3}$	17/14/18	15/13/17	11/7.9/15	10/8.4/12	11/9.3/12	7.4/6.4/8.8	7.8/7.4/8.2	13/12/15	59/45/76	24/20/28
$64, 2\times10^{-3}$	20/20/20	18/18/19	9.5/8.4/11	10/9.1/10	**9.6/9.4/11**	16/11/21	21/18/25	27/23/28	79/70/91	40/38/42
$128, 2\times10^{-3}$	20/18/22	18/16/21	9.7/8.8/11	11/10/13	11/11/12	9.7/6.3/14	14/9.3/20	20/15/24	79/68/99	33/26/42
$256, 2\times10^{-3}$	20/19/21	19/18/19	8.6/8.1/8.9	13/11/15	13/12/15	13/9.8/18	25/18/36	34/26/43	84/75/92	42/34/48
$512, 2\times10^{-3}$	18/13/20	16/12/20	43/30/54	59/40/72	62/45/81	28/19/39	18/14/20	23/8/31	70/42/99	35/27/44
$32, 4\times10^{-4}$	19/17/21	17/15/19	11/10/12	13/11/17	13/11/14	6.6/4.9/8.6	6.9/6.5/7.5	12/11/12	57/43/77	22/18/25
$64, 4\times10^{-4}$	16/13/18	14/11/16	8.9/7.8/10	12/6.8/18	13/8.5/19	12/6.7/22	17/6.2/29	21/9.1/32	75/48/93	33/20/43
$128, 4\times10^{-4}$	17/17/17	16/16/16	10/8.8/12	11/10/12	13/9.7/17	14/8.7/20	24/15/32	29/21/36	82/82/83	39/33/45
$256, 4\times10^{-4}$	21/13/27	21/13/27	11/11/11	14/11/16	21/14/28	44/28/56	59/40/83	71/43/100	*111/53/158*	*72/47/100*
$512, 4\times10^{-4}$	19/18/21	18/17/19	17/11/23	25/18/30	30/23/35	40/16/73	*59/32/100*	*69/41/110*	103/98/109	68/50/96

TABLE C1. Scores obtained by each configuration for our metrics panel. Reported scores correspond to the best score obtained after training saturation for the 3 runs, in the order average/best/worst. For all metrics considered, lower is better. For a given configuration and and a given run, all "best scores after saturation" do not necessarily correspond to the same step for all metrics. Overall best scores in bold black, worst scores in italic.

Estimator	$s_{21,u}$	$s_{21,v}$	$s_{21,t_{2m}}$	$s_{22,u}$	$s_{22,v}$	$s_{22,t_{2m}}$
Scale	$\times 10^3$			$\times 10^2$		
$32, 4\times10^{-3}$	5.0	3.8	5.6	4.7	4.6	1.7
$64, 4\times10^{-3}$	5.0	4.3	5.0	5.4	4.8	2.6
$256, 4\times10^{-3}$	5.0	4.7	6.8	7.7	5.5	5.2
$32, 2\times10^{-3}$	**2.7**	**1.9**	4.0	**3.4**	**2.4**	**1.3**
$64, 2\times10^{-3}$	3.4	**1.8**	**3.4**	4.4	3.9	1.5

TABLE C2. Scattering RMSE estimators, for each variable and average over the 3 runs. Best results in bold. Shown are only the most successful configurations as results from Table C1. The score ranking previously obtained with PSD and SWD slightly changes with this metric.

An important consistency check consists in verifying that the GAN does not memorize the training samples, and is able to generate sufficiently different samples. To examine this aspect, we use the Mean Square Error (MSE) and look for the pair of GAN and AROME-EPS samples with the lowest global (i.e including all variables and grid points) MSE distance, across both GAN-generated and AROME-EPS datasets. Such samples are shown on Figure D1, and exhibit noticeable visual difference. The distribution of global MSE distance of this specific GAN sample

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

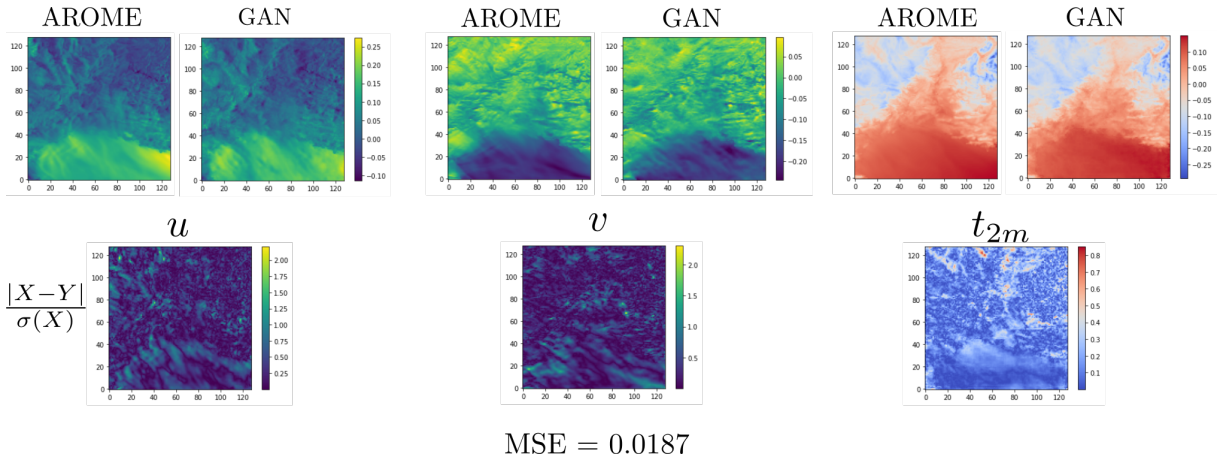

FIG. D1. Comparing the MSE-nearest samples in the GAN and AROME-EPS datasets. These samples clearly differ from each other. In the bottom row plot, the pixel-wise absolute distance is compared to the pixel-wise standard deviation of the AROME-EPS sample. The global MSE related to this sample is reported at the bottom.

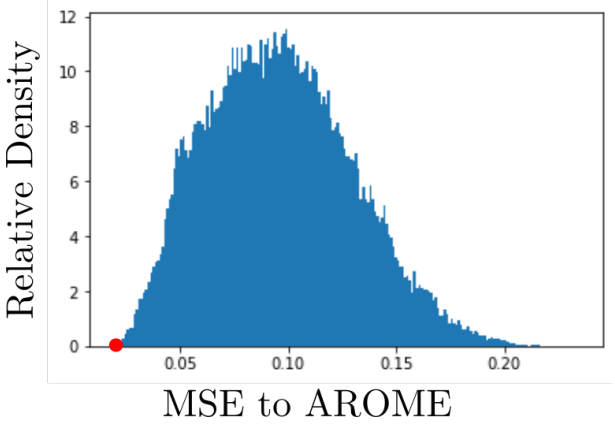

FIG. D2. Visualizing the distribution of MSE distance from AROME-EPS to the GAN sample of Figure D1. The red dot denotes the distance to the nearest AROME-EPS sample. The distribution is broad and its peak is of the order of magnitude of the normalized dataset variance.

with the whole AROME-EPS dataset is also plotted. As can be seen on Figure D2, this distribution peaks at about 0.1, which is approximately the variance of the normalized AROME-EPS dataset. The MSE-minimum is thus sensibly distinct from any AROME sample, while being at a consistent average MSE distance from the whole dataset, further confirming the absence of mode collapse.

Accepted for publication in *Artificial Intelligence for the Earth Systems*. DOI 10.1175/AIES-D-23-0006.1.

References

Alsallakh, B., N. Kokhlikyan, V. Miglani, J. Yuan, and O. Reblitz-Richardson, 2021: Mind the pad – {cnn}s can develop blind spots. *International Conference on Learning Representations*.

Andreux, M., and Coauthors, 2018: Kymatio: Scattering transforms in python. arXiv, https://doi.org/10.48550/ARXIV.1812.11214.

Andén, J., and S. Mallat, 2014: Deep scattering spectrum. *IEEE Transactions on Signal Processing*, **62 (16)**, 4114–4128, https://doi.org/10.1109/TSP.2014.2326991.

Arjovsky, M., S. Chintala, and L. Bottou, 2017: Wasserstein generative adversarial networks. *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, JMLR.org, 214–223.

Bengio, Y., and X. Glorot, 2010: Understanding the difficulty of training deep feed forward neural networks. *International Conference on Artificial Intelligence and Statistics*, 249–256.

Besombes, C., O. Pannekoucke, C. Lapeyre, B. Sanderson, and O. Thual, 2021: Producing realistic climate data with generative adversarial networks. *Nonlinear Processes in Geophysics*, **28 (3)**, 347–370, https://doi.org/10.5194/npg-28-347-2021.

Bhatia, S., A. Jain, and B. Hooi, 2021: Exgan: Adversarial generation of extreme samples. 2009.08454.

Bihlo, A., 2020: A generative adversarial network approach to (ensemble) weather prediction. 2006.07718.

Blau, Y., and T. Michaeli, 2018: The perception-distortion tradeoff. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6228–6237, https://doi.org/10.1109/CVPR.2018.00652.

Bouttier, F., L. Raynaud, O. Nuissier, and B. Ménétrier, 2016: Sensitivity of the arome ensemble to initial and surface perturbations during hymex. *Quarterly Journal of the Royal Meteorological Society*, **142 (S1)**, 390–403, https://doi.org/https://doi.org/10.1002/qj.2622, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.2622.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

Bouttier, F., B. Vié, O. Nuissier, and L. Raynaud, 2012: Impact of stochastic physics in a convection-permitting ensemble. *Monthly Weather Review*, **140 (11)**, 3706 – 3721, https://doi.org/10.1175/MWR-D-12-00031.1.

Brock, A., J. Donahue, and K. Simonyan, 2018: Large scale GAN training for high fidelity natural image synthesis. *CoRR*, **abs/1809.11096**, 1809.11096.

Brousseau, P., L. Berre, F. Bouttier, and G. Desroziers, 2011: Background-error covariances for a convective-scale data-assimilation system: Arome–france 3d-var. *Quarterly Journal of the Royal Meteorological Society*, **137 (655)**, 409–422, https://doi.org/https://doi.org/10.1002/qj.750, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.750.

Bruna, J., and S. Mallat, 2013: Invariant scattering convolution networks. *IEEE transactions on pattern analysis and machine intelligence*, **35**, 1872–1886, https://doi.org/10.1109/TPAMI.2012.230.

Cheng, S., and B. Ménard, 2021: How to quantify fields or textures? a guide to the scattering transform. arXiv, https://doi.org/10.48550/ARXIV.2112.01288.

Cheng, S., Y.-S. Ting, B. Ménard, and J. Bruna, 2020: A new approach to observational cosmology using the scattering transform. *Monthly Notices of the Royal Astronomical Society*, **499 (4)**, 5902–5914, https://doi.org/10.1093/mnras/staa3165, https://academic.oup.com/mnras/article-pdf/499/4/5902/34157889/staa3165.pdf.

Denis, B., J. Côté, and R. Laprise, 2002: Spectral decomposition of two-dimensional atmospheric fields on limited-area domains using the discrete cosine transform (dct). *Monthly Weather Review*, **130 (7)**, 1812 – 1829, https://doi.org/10.1175/1520-0493(2002)130⟨1812:SDOTDA⟩2.0.CO;2.

Descamps, L., C. Labadie, A. Joly, E. Bazile, P. Arbogast, and P. Cébron, 2015: Pearp, the météo-france short-range ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, **141 (690)**, 1671–1685, https://doi.org/https://doi.org/10.1002/qj.2469, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.2469.

Dumoulin, V., I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville, 2016: Adversarially learned inference. arXiv, https://doi.org/10.48550/ARXIV.1606.00704.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

Ebert, E. E., 2008: Fuzzy verification of high-resolution gridded forecasts: a review and proposed framework. *Meteorological Applications*, **15 (1)**, 51–64, https://doi.org/https://doi.org/10.1002/met.25, URL https://rmets.onlinelibrary.wiley.com/doi/abs/10.1002/met.25, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/met.25.

Gagne II, D. J., H. M. Christensen, A. C. Subramanian, and A. H. Monahan, 2020: Machine learning for stochastic parameterization: Generative adversarial networks in the lorenz '96 model. *Journal of Advances in Modeling Earth Systems*, **12 (3)**, e2019MS001 896, https://doi.org/https://doi.org/10.1029/2019MS001896, URL https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/2019MS001896, e2019MS001896 10.1029/2019MS001896, https://agupubs.onlinelibrary.wiley.com/doi/pdf/10.1029/2019MS001896.

Garcia, G. B., M. Lagrange, I. Emmanuel, and H. Andrieu, 2015: Classification of rainfall radar images using the scattering transform. *2015 23rd European Signal Processing Conference (EUSIPCO)*, 1940–1944, https://doi.org/10.1109/EUSIPCO.2015.7362722.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, 2014: Generative adversarial networks. arXiv, https://doi.org/10.48550/ARXIV.1406.2661.

Gulrajani, I., F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville, 2017: Improved training of wasserstein gans. *CoRR*, **abs/1704.00028**, 1704.00028.

Harris, L., A. T. T. McRae, M. Chantry, P. D. Dueben, and T. N. Palmer, 2022: A generative deep learning approach to stochastic downscaling of precipitation forecasts. arXiv, https://doi.org/10.48550/ARXIV.2204.02028.

Karras, T., T. Aila, S. Laine, and J. Lehtinen, 2018: Progressive growing of gans for improved quality, stability, and variation. 1710.10196.

Karras, T., S. Laine, and T. Aila, 2019: A style-based generator architecture for generative adversarial networks. 1812.04948.

Karras, T., S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, 2020: Analyzing and improving the image quality of stylegan. 1912.04958.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

Kingma, D., and M. Welling, 2014: Auto-encoding variational bayes.

Kingma, D. P., and J. Ba, 2015: Adam: A method for stochastic optimization. *CoRR*, **abs/1412.6980**.

Kingma, D. P., and M. Welling, 2019: An introduction to variational autoencoders. *Foundations and Trends® in Machine Learning*, **12 (4)**, 307–392, https://doi.org/10.1561/2200000056.

Kolouri, S., P. E. Pope, C. E. Martin, and G. K. Rohde, 2018: Sliced-wasserstein autoencoder: An embarrassingly simple generative model. 1804.01947.

Kynkäänniemi, T., T. Karras, S. Laine, J. Lehtinen, and T. Aila, 2019: Improved precision and recall metric for assessing generative models. 1904.06991.

Leinonen, J., D. Nerini, and A. Berne, 2021: Stochastic super-resolution for downscaling time-evolving atmospheric fields with a generative adversarial network. *IEEE Transactions on Geoscience and Remote Sensing*, **59 (9)**, 7211–7223, https://doi.org/10.1109/tgrs.2020.3032790.

Lim, J. H., and J. C. Ye, 2017: Geometric gan. 1705.02894.

Mallat, S., 2012: Group invariant scattering. *Communications on Pure and Applied Mathematics*, **65 (10)**, 1331–1398, https://doi.org/https://doi.org/10.1002/cpa.21413, https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpa.21413.

Marin, I., S. Gotovac, M. Russo, and D. Božić-Štulić, 2021: The effect of latent space dimension on the quality of synthesized human face images. *Journal of communications software and systems*, **17 (2)**, 124–133, https://doi.org/10.24138/jcomss-2021-0035.

Mescheder, L., A. Geiger, and S. Nowozin, 2018: Which training methods for gans do actually converge? https://doi.org/10.48550/ARXIV.1801.04406.

Miyato, T., T. Kataoka, M. Koyama, and Y. Yoshida, 2018: Spectral normalization for generative adversarial networks. *CoRR*, **abs/1802.05957**, 1802.05957.

Montmerle, T., Y. Michel, E. Arbogast, B. Ménétrier, and P. Brousseau, 2018: A 3d ensemble variational data assimilation scheme for the limited-area arome model: Formulation and

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

preliminary results. *Quarterly Journal of the Royal Meteorological Society*, **144 (716)**, 2196–2215, https://doi.org/https://doi.org/10.1002/qj.3334, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.3334.

Mustafa, M., D. Bard, W. Bhimji, Z. Lukić, R. Al-Rfou, and J. M. Kratochvil, 2019: CosmoGAN: creating high-fidelity weak lensing convergence maps using generative adversarial networks. *Computational Astrophysics and Cosmology*, **6 (1)**, https://doi.org/10.1186/s40668-019-0029-9, URL https://doi.org/10.1186%2Fs40668-019-0029-9.

Odena, A., C. Olah, and J. Shlens, 2017: Conditional image synthesis with auxiliary classifier gans. 1610.09585.

Olea, R. A., 1994: Fundamentals of Semivariogram Estimation, Modeling, and Usage. *Stochastic Modeling and Geostatistics: Principles, Methods, and Case Studies*, American Association of Petroleum Geologists, https://doi.org/10.1306/CA3590C4.

Pannekoucke, O., L. Berre, and G. Desroziers, 2008: Background-error correlation length-scale estimates and their sampling statistics. *Quarterly Journal of the Royal Meteorological Society*, **134 (631)**, 497–508, https://doi.org/https://doi.org/10.1002/qj.212, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.212.

Pantillon, F., P. Knippertz, and U. Corsmeier, 2017: Revisiting the synoptic-scale predictability of severe european winter storms using ecmwf ensemble reforecasts. *Natural Hazards and Earth System Sciences*, **17 (10)**, 1795–1810, https://doi.org/10.5194/nhess-17-1795-2017.

Paszke, A., and Coauthors, 2019: *PyTorch: An Imperative Style, High-Performance Deep Learning Library*. Curran Associates Inc., Red Hook, NY, USA.

Ponzano, M., B. Joly, L. Descamps, and P. Arbogast, 2020: Systematic error analysis of heavy-precipitation-event prediction using a 30-year hindcast dataset. *Natural Hazards and Earth System Sciences*, **20 (5)**, 1369–1389, https://doi.org/10.5194/nhess-20-1369-2020.

Rabin, J., G. Peyré, J. Delon, and B. Marc, 2011: Wasserstein Barycenter and its Application to Texture Mixing. *SSVM'11*, Springer, Israel, 435–446.

Radford, A., L. Metz, and S. Chintala, 2015: Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv e-prints*, arXiv:1511.06434, 1511.06434.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

Ramdas, A., N. Garcia, and M. Cuturi, 2015: On wasserstein two sample testing and related families of nonparametric tests. arXiv, https://doi.org/10.48550/ARXIV.1509.02237.

Ravuri, S., and Coauthors, 2021: Skilful precipitation nowcasting using deep generative models of radar. *Nature*, **597, 672–677 (2021)**, 672–677, https://doi.org/10.1038/s41586-021-03854-z.

Raynaud, L., and F. Bouttier, 2016: Comparison of initial perturbation methods for ensemble prediction at convective scale. *Quarterly Journal of the Royal Meteorological Society*, **142 (695)**, 854–866, https://doi.org/https://doi.org/10.1002/qj.2686, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.2686.

Raynaud, L., and O. Pannekoucke, 2013: Sampling properties and spatial filtering of ensemble background-error length-scales. *Quarterly Journal of the Royal Meteorological Society*, **139 (672)**, 784–794, https://doi.org/https://doi.org/10.1002/qj.1999, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.1999.

Roberts, N. M., and H. W. Lean, 2008: Scale-selective verification of rainfall accumulations from high-resolution forecasts of convective events. *Monthly Weather Review*, **136 (1)**, 78 – 97, https://doi.org/https://doi.org/10.1175/2007MWR2123.1, URL https://journals.ametsoc.org/view/journals/mwre/136/1/2007mwr2123.1.xml.

Rubner, Y., C. Tomasi, and L. J. Guibas, 2004: The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, **40**, 99–121.

Saxe, A. M., J. L. McClelland, and S. Ganguli, 2014: Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. 1312.6120.

Seity, Y., P. Brousseau, S. Malardel, G. Hello, P. Bénard, F. Bouttier, C. Lac, and V. Masson, 2011: The arome-france convective-scale operational model. *Monthly Weather Review*, **139 (3)**, 976 – 991, https://doi.org/10.1175/2010MWR3425.1.

Sergeev, A., and M. D. Balso, 2018: Horovod: fast and easy distributed deep learning in tensorflow. *CoRR*, **abs/1802.05799**, 1802.05799.

Sha, Y., D. J. G. II, G. West, and R. Stull, 2020: Deep-learning-based gridded downscaling of surface meteorological variables in complex terrain. part ii: Daily precipitation.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.

Journal of Applied Meteorology and Climatology, **59 (12)**, 2075 – 2092, https://doi.org/10.1175/JAMC-D-20-0058.1.

Vincendon, B., V. Ducrocq, O. Nuissier, and B. Vié, 2011: Perturbation of convection-permitting nwp forecasts for flash-flood ensemble forecasting. *Natural Hazards and Earth System Sciences*, **11 (5)**, 1529–1544, https://doi.org/10.5194/nhess-11-1529-2011, URL https://nhess.copernicus.org/articles/11/1529/2011/.

Weaver, A. T., and I. Mirouze, 2013: On the diffusion equation and its application to isotropic and anisotropic correlation modelling in variational assimilation. *Quarterly Journal of the Royal Meteorological Society*, **139 (670)**, 242–260, https://doi.org/https://doi.org/10.1002/qj.1955, https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.1955.

Xu, R., X. Wang, K. Chen, B. Zhou, and C. C. Loy, 2020: Positional encoding as spatial inductive bias in gans. arXiv, https://doi.org/10.48550/ARXIV.2012.05217.

Zhang, H., I. Goodfellow, D. Metaxas, and A. Odena, 2019: Self-attention generative adversarial networks. 1805.08318.

Zhang, R., 2019: Making convolutional networks shift-invariant again. *ICML*.

Accepted for publication in *Artificial Intelligence for the Earth Systems.* DOI 10.1175/AIES-D-23-0006.1.