Discussion of "Saving Storage in Climate Ensembles: A Model-Based Stochastic Approach"

Abhirup DATTA

Huang et al (J Agric Biol Environ Stat, 2023, https://doi.org/10.1007/s13253-022-00518-x) a suite of statistical models for storage-efficient climate model emulation. In this discussion, I review and explore possibility of using machine learning methods, in particular, deep neural network (DNN)-based variational autoencoders (VAE) for the same task of spatio-temporal climate data compression. I discuss the pros and cons of the statistical and the machine learning paradigms.

Key Words: Machine learning; Deep neural network; Variational autoencoders; Climate data compression.

I congratulate the authors of Huang et al. (2023) on a very interesting paper. Stochastic emulation of climate models has been studied for many decades now. However, the primary objective has often been to develop fast statistical models that offer approximate output at a fraction of the computational time needed for actual climate model runs. This manuscript shifts the focus on another important resource—data storage, that is becoming increasingly expensive as the size and number of datasets grow rapidly. The authors develop a suite of spatio-temporal models to emulate climate model outputs at increasing spatio-temporal resolutions—starting with global annual data, and ending at daily data on a fine spatial grid over the surface of the earth. Consequently, the statistical models proposed to emulate these outputs also become progressively more complex.

For each data resolution, the authors propose parsimonious yet sufficiently rich model classes to adequately represent the data while regulating the parameter dimensionality. Examples of such well-motivated model choices include adding harmonics for intra-annual variation in temperature, anisotropic modeling of spatial dependence on the surface of the earth via first modeling the spatial processes along each latitude (via axial symmetry) and correlating these latitude-specific random fields via a coherence model. Despite these

This article is a commentary for https://doi.org/10.1007/s13253-022-00518-x.

A. Datta (✉) Department of Biostatistics, Johns Hopkins University, Baltimore, USA
(E-mail: *abhidatta@jhu.edu*).

Journal of Agricultural, Biological, and Environmental Statistics, Volume 28, Number 2, Pages 352–357
https://doi.org/10.1007/s13253-023-00539-0

efforts to preserve parsimony, the most complex models considered in the paper have nearly a million parameters. This is not an issue from a storage perspective as the storage cost of saving a million numbers is negligible, relative to the cost of saving the entire outputs. This raises the question whether machine learning models of similar dimensionality can also be worthy candidates for the same task. In this discussion, I review and explore use of *deep neural network* (DNN)-based *variational autoencoders* (VAE) (Kingma and Welling 2014) for the task of spatio-temporal climate data compression.

VARIATIONAL AUTOENCODERS (VAE)

I first briefly summarize the core ideas of VAE here that will be needed for the discussion. Girin et al. (2020) offers a more detailed statistical review of VAE for temporal data. VAE are unsupervised learners using variational Bayes that model the observed high-dimensional data $\boldsymbol{x} \in \mathbb{R}^D$ in terms of a much lower-dimensional latent variable $\boldsymbol{z} \in \mathbb{R}^L$, $(L \ll D)$. The latent variable $\boldsymbol{z}$ is commonly assigned a standard normal prior, i.e., $\boldsymbol{z} \sim N(\boldsymbol{0}, \boldsymbol{I}_L)$. This assumption is reasonable as latent Gaussian random effects are abundant in spatio-temporal mixed models. VAE then specify two distributions—the data distribution or *generative distribution* $p(\boldsymbol{x} \mid \boldsymbol{z}; \boldsymbol{\theta}_x)$ and the *variational* distribution $q(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\theta}_z)$. Variational Bayes thus differs from standard Bayesian inference in introducing this variational distribution, which can be thought of as an analytically tractable surrogate for the posterior distribution $p(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\theta}_x)$. In fact, if the variational family $q(\cdot)$ is unconstrained, the posterior distribution is indeed the optimal choice for it. In practice, $q(\cdot)$ is constrained to be from a tractable family of distributions like Gaussian, i.e.,

$$q(\boldsymbol{z} \mid \boldsymbol{x}; \boldsymbol{\theta}_z) = N\left(\boldsymbol{z} \mid \boldsymbol{\mu}_z(\boldsymbol{x}; \boldsymbol{\theta}_z), \operatorname{diag}(\boldsymbol{\sigma}_z^2(\boldsymbol{x}; \boldsymbol{\theta}_z))\right).$$

This is a pragmatic choice to facilitate fast computation. To allow flexibility within the chosen variational family, the means $\boldsymbol{\mu}_z$ and variances $\boldsymbol{\sigma}_z^2$ are modeled using a rich family of functions, *deep neural networks* (DNN) (LeCun et al. 2015). A K-layer neural network is a function class expressed recursively as $f_{\boldsymbol{\theta}}(\boldsymbol{x}) = g_K(\boldsymbol{a}_k + \boldsymbol{W}_k g_{K-1}(\boldsymbol{a}_{K-1} + \boldsymbol{W}_{K-1} g_{K-2}(\ldots g_1(\boldsymbol{a}_1 + \boldsymbol{W}_1 \boldsymbol{x}) \ldots))$ where g_k's are known *activation functions* (links) and $\boldsymbol{\theta}$ is the set of all parameters comprising of the weights $\boldsymbol{W}_k$ and biases $\boldsymbol{a}_k$, for $k = 1, \ldots, K$. The variational family $q(\cdot)$ with means $\boldsymbol{\mu}_z$ and variances $\boldsymbol{\sigma}_z^2$ modeled as DNN is known as the *encoder* network. The data distribution $p(\boldsymbol{x} \mid \boldsymbol{z}; \boldsymbol{\theta}_x)$ is also modeled using DNN and is referred to as the *decoder* network. For continuous data, a common choice is the Gaussian likelihood $p(\boldsymbol{x} \mid \boldsymbol{z}; \boldsymbol{\theta}_x) = N(\boldsymbol{x} \mid \boldsymbol{\mu}_x(\boldsymbol{z}; \boldsymbol{\theta}_x), \operatorname{diag}(\boldsymbol{\sigma}_x^2(\boldsymbol{z}; \boldsymbol{\theta}_x)))$ with means and variances modeled as DNN.

In variational inference, the *generative* parameters $\boldsymbol{\theta}_x$ and the *variational* parameters $\boldsymbol{\theta}_z$ are obtained by maximizing the *variational lower bound* (VLB) or *evidence lower bound* (ELBO) (Neal and Hinton 1998)

$$\mathcal{L}(\boldsymbol{x}; \boldsymbol{\theta}_x, \boldsymbol{\theta}_z) = E_{\boldsymbol{z} \sim q} \log p(\boldsymbol{x} \mid \boldsymbol{z}; \boldsymbol{\theta}_x) - D_{KL}(q(\boldsymbol{z}; \boldsymbol{\theta}_z) \| p(\boldsymbol{z})) \tag{1}$$

where D_{KL} denotes the Kullback–Leibler divergence. The optimization is typically conducted using a *stochastic gradient descent* (SGD) algorithm (Kingma and Ba 2014).

VAE FOR COMPRESSING CLIMATE ENSEMBLES

Autoencoders are ideal for the task of data compression as they are essentially dimension reduction techniques. The high-dimensional data $\boldsymbol{x}$ are compressed (encoded) to a low-dimensional representation $\boldsymbol{z}$ which can be used to reconstruct (decode) $\hat{\boldsymbol{x}}$ (an approximation of $\boldsymbol{x}$). The latent embeddings $\boldsymbol{z}$ can be viewed as nonlinear analogs of principal components (PC) often referred to as *empirical orthogonal functions* (EOFs). VAE often offer a better representation of the data than EOF in atmospheric applications (Krinitskiy et al. 2019). Consequently, DNN and VAE are being increasingly adopted in atmospheric and climate applications. Rasp et al. (2018) used DNN-based emulators of subgrid processes. Behrens et al. (2022) achieved nearly similar performance for the same task while using a DNN-VAE with a very low-dimensional latent space. Cartwright et al. (2021) used convolutional neural network (CNN)-based VAE for emulating gridded data on gas plumes. Saenz et al. (2018) considered a similar application to Huang et al. (2023) in compressing surface temperature output from climate models and used CNN-based autoencoders for dimension-reduction of gridded temperature data.

None of the aforementioned applications consider data from an ensemble of outputs, and the goal was to primarily to reconstruct one dataset (Saenz et al. 2018) and conduct additional tasks like spatio-temporal intrapolation (Cartwright et al. 2021) or interpretation of the latent variable (Behrens et al. 2022) (akin to interpreting principal components). The goal of Huang et al. (2023) is different—it is to compress the output with the aim of being able to reconstruct not just one dataset (or each individual dataset from the ensemble used in the training) but the underlying distribution generating the ensemble of datasets. VAE are particularly suited for this task of compressing climate ensembles by only storing distributional parameters (referred to as *unconditional compression* in the paper). I outline a simple generic algorithm below.

Consider a climate model output consisting of N ensembles $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N$. For now, I make no assumptions about the nature of the outputs $\boldsymbol{x}_r$. They can be univariate or multivariate time-series (as in the global and regional outputs considered in the paper) or a spatio-temporal lattice (for the gridded outputs). These data-types will dictate the specific architecture of the DNN-VAE, but the generic algorithm remains same for all of them. To generate an ensemble of outputs from the same population as the observed data, one needs to learn the underlying distribution $\mathcal{F}_x$ generating this data. The VAE models this data distribution using a *Bayesian hierarchical model* (BHM),

$$p(\boldsymbol{x}_r; \boldsymbol{\theta}) = \int p(\boldsymbol{x}_r \mid z_r; \boldsymbol{\theta}_x) p(z_r) dz_r \tag{2}$$

where z_r is the latent variable corresponding to the rth member of the ensemble. As discussed before, $p(z_r)$ will often be the canonical Gaussian distribution, $p(z_r) = N(z_r \mid \mathbf{0}, \boldsymbol{I}_L)$ while $p(\boldsymbol{x}_r \mid z_r; \boldsymbol{\theta}_x)$ will depend on the data type. For continuous data, one will often

use Gaussian likelihood $p(\boldsymbol{x}_r \mid z_r; \boldsymbol{\theta}_x) = N(\boldsymbol{x}_r \mid \boldsymbol{\mu}_x(z_r; \boldsymbol{\theta}_x), \text{diag}(\boldsymbol{\sigma}_x^2(z_r; \boldsymbol{\theta}_x)))$. Here, $\boldsymbol{\mu}_x()$ and $\boldsymbol{\sigma}_x^2()$ are functions in the DNN class. Thus, to learn the *generative distribution* $\mathcal{F}_x$, one needs to estimate the combined set of parameters $\boldsymbol{\theta}_x$ for these mean and variance DNN. However, the marginal likelihood in (2) will be often intractable, and even sampling from the posterior $p(z_r \mid \boldsymbol{x}_r)$ will be challenging due to nonlinearity of DNN. Hence, VAE deploys the variational family $q(z_r) = N(z_r \mid \boldsymbol{\mu}_{z,r}, \text{diag}(\boldsymbol{\sigma}_{z,r}^2))$ where the variational means $\boldsymbol{\mu}_{z,r} = \boldsymbol{\mu}_z(\boldsymbol{x}_r, \boldsymbol{\theta}_z)$ and variances $\boldsymbol{\sigma}_{z,r}^2 = \boldsymbol{\sigma}_z^2(\boldsymbol{x}_r, \boldsymbol{\theta}_z)$ are modeled as DNN. The parameters $(\boldsymbol{\theta}_x, \boldsymbol{\theta}_z)$ can be estimated by optimizing the total ELBO (1) over the ensemble, i.e., $\sum_{r=1}^N \mathcal{L}(\boldsymbol{x}_r \mid \boldsymbol{\theta}_x, \boldsymbol{\theta}_z)$.

When autoencoders are used for reconstruction tasks, one needs to store all the lower-dimensional z_r's (or their variational means $\boldsymbol{\mu}_{z,r}$'s). For unconditional compression, the goal is not to reconstruct one dataset but to estimate the generative distribution. Hence, one only needs to store estimates of the generative model parameters $\boldsymbol{\theta}_x$. Storage of the z_r or their variational means and variances are not required. Thus the parameters of the variational family $\boldsymbol{\theta}_z$ are simply nuisance parameters for this task. Conditional on the knowledge of the generative parameters $\boldsymbol{\theta}_x$, one can simulate from $\mathcal{F}_x$ using the hierarchical formulation in (2), i.e., simulating $z_{new} \sim N(\boldsymbol{0}, \boldsymbol{I}_L)$ and then simulating $\boldsymbol{x}_{new}$ from $p(\boldsymbol{x}_{new} \mid z_{new}, \boldsymbol{\theta}_x)$. This allows to regenerate a new ensemble $\boldsymbol{x}_{new,1}, \boldsymbol{x}_{new,2}, \ldots$ from the distribution $\mathcal{F}_x$.

The architecture of the DNN used in the encoder and the decoder will depend on the nature of the data. For the global and regional models considered in the paper, each $\boldsymbol{x}_r$ is a univariate or multivariate time-series of surface temperatures. One would then consider using variational *recurrent neural networks* (RNN) (Chung et al. 2015) or some variants like *long short-term memory* (LSTM) networks, *echo state networks* (ESN) that can model dependencies over time. For the spatio-temporal gridded data, RNN and CNN can be combined (Wang et al. 2016). CNN are designed for image valued data and are ideal to model data observed over a spatial lattice. Hence, CNN has already seen considerable use for compressing gridded climate output (Saenz et al. 2018; Cartwright et al. 2021). Kernels with horizontal or vertical contours can be used as filters in the CNN to capture variation along the latitudes and longitudes, respectively. The temporal evolution can still be captured by an RNN variant. Additional covariate information like radioactive forcing or land/ocean classification data can be incorporated by specifying the variation family as $q(z_r \mid \boldsymbol{x}_r^*, \boldsymbol{\theta}_z)$ where $\boldsymbol{x}_r^*$ is the augmented data containing $\boldsymbol{x}_r$ and these covariates. Finally, if Gaussianity is inappropriate, as in the daily gridded data, one can easily switch to a different generative model $p(\boldsymbol{x}_r \mid z_r, \boldsymbol{\theta}_z)$ like the Tukey family of distributions considered in the paper.

Validation of the trained VAE for this task would also need to be different from the usual validation of VAE that assesses reconstruction accuracy on the test data. Instead using the generated ensemble, one would need to assess the distributional properties using metrics similar to the ones used in Huang et al. (2023).

CONCLUDING REMARKS

DNN and other deep models are increasingly becoming mainstream in spatio-temporal analysis (see Wikle and Zammit-Mangion 2022, for a comprehensive review). In this discus-

sion, I proposed considering DNN-based VAE as an alternative to the statistical models of Huang et al. (2023) for storage-efficient emulation of climate ensembles. I end the discussion with a comparison of the two paradigms with respect to three considerations.

Storage: Like the statistical models considered, the VAE approach would also need to only store estimates of the generative model parameters. This parameter set for VAE will be considerably larger than many of the models proposed in the paper. However, as demonstrated in the paper, storing even a million parameter estimates only uses a negligible fraction of the memory required for storing the entire ensemble; hence, the complexity of the DNN-VAE models should not be a concern from a storage perspective.

Interpretation: In the paper, the authors have done a commendable job of keeping the models parsimonious, and motivating every added complexity to the extent possible. In deep models, parsimony is still valued and is enforced through past-dependence in RNN or via convolution and pooling in CNN. Despite this, these models are generally much less and more black-box (or descriptive). However, deep models have been shown to retain some interpretability in climate applications. For example, the DNN emulators used in Rasp et al. (2018) were shown to learn to conserve energy just based on the training. Behrens et al. (2022) demonstrated how the latent embedding from VAE can be used to identify dominant drivers of convective processes. Thus, deep models can learn and represent features of the data without needing explicit coding of these in the model architecture.

Accuracy: The richer representation of underlying processes offered by deep models may manifest in improved accuracy of reconstruction over parametric statistical models. However, one downside of the deep models is that they require larger amounts of training data. Saenz et al. (2018) demonstrated that with small training data, EOF outperform CNN-based VAE for compressing surface temperature output, but the trend reverses when using a larger training data. Thus, for smaller datasets using parametric spatio-temporal models or hybrid (semi-parametric) statistical-machine learning models (Wikle and Zammit-Mangion 2022; Saha et al. 2021; Sigrist 2020) might be more suitable.

In the future, detailed empirical comparisons need to be undertaken to assess benefits and pitfalls of statistical, machine learning and hybrid approaches for compressing climate ensembles.

[Accepted March 2023. Published Online May 2023.]

REFERENCES

Behrens G, Beucler T, Gentine P, Iglesias-Suarez F, Pritchard M, Eyring V (2022) Non-linear dimensionality reduction with a variational encoder decoder to understand convective processes in climate models. J Adv Model Earth Syst 14(8):e2022MS003130

Cartwright L, Zammit-Mangion A, Deutscher NM (2021) Emulation of greenhouse-gas sensitivities using variational autoencoders. arXiv preprint arXiv:2112.12524

Chung J, Kastner K, Dinh L, Goel K, Courville AC, Bengio Y (2015) A recurrent latent variable model for sequential data. Adv Neural Inf Process Syst, *28*

Girin L, Leglaive S, Bie X, Diard J, Hueber T, Alameda-Pineda X (2020) Dynamical variational autoencoders: a comprehensive review. arXiv preprint arXiv:2008.12595

Huang H, Castruccio S, Baker AH, Genton MG (2023) Saving storage in climate ensembles: a model-based stochastic approach. J Agric Biol Environ Stat. https://doi.org/10.1007/s13253-022-00518-x

Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980

Kingma DP, Welling M (2014) Auto-encoding variational bayes. In: International conference on learning representations

Krinitskiy MA, Zyulyaeva YA, Gulev SK (2019) Clustering of polar vortex states using convolutional autoencoders. CEUR Workshop Proceedings 2426:52–61

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444

Neal RM, Hinton GE (1998) A view of the EM algorithm that justifies incremental, sparse, and other variants. In: Learning in graphical models, Springer, pp 355–368

Rasp S, Pritchard MS, Gentine P (2018) Deep learning to represent subgrid processes in climate models. Proc Natl Acad Sci 115(39):9684–9689

Saenz JA, Lubbers N, Urban NM (2018) Dimensionality-reduction of climate data using deep autoencoders. arXiv preprint arXiv:1809.00027

Saha A, Basu S, Datta A (2021) Random forests for spatially dependent data. J Am Stat Assoc 118(541):665–683. https://doi.org/10.1080/01621459.2021.1950003

Sigrist F (2020) Gaussian process boosting. arXiv preprint arXiv:2004.02653

Wang J, Yang Y, Mao J, Huang Z, Huang C, Xu W (2016) Cnn-rnn: a unified framework for multi-label image classification. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2285–2294

Wikle CK, Zammit-Mangion A (2022) Statistical deep learning for spatial and spatio-temporal data. arXiv preprint arXiv:2206.02218