TESTING INTERPRETABILITY TECHNIQUES FOR DEEP STATISTICAL CLIMATE DOWNSCALING

Jose González-Abad
Meteorology Group
Advanced Computing and e-Science Group
Instituto de Fısica de Cantabria (IFCA),
CSIC- Universidad de Cantabria
Santander, Spain
gonzabad@ifca.unican.es

Jorge Baño-Medina & José Manuel Gutiérrez
Meteorology Group
Instituto de Fısica de Cantabria (IFCA),
CSIC- Universidad de Cantabria
Santander, Spain

ABSTRACT

Deep Learning (DL) has recently emerged as a promising Empirical Statistical Downscaling *perfect-prognosis* technique (ESD-PP), to generate high-resolution fields from large-scale climate variables. Here, we analyze two state-of-the-art DL topologies for ESD-PP of different levels of complexity over North America. Besides classical validation leaning on accuracy metrics (e.g., Root Mean Squared Error (RMSE)), we evaluate several interpretability techniques to gain understanding on the inner functioning of the DL models deployed. Taking as reference the RMSE both topologies show similar values. Nonetheless, by analyzing the resulting interpretability maps, we find that the simplest model fails to capture a realistic physics-based input-output link, whilst the complex one describes a local pattern, characteristic of downscaling. In climate change scenarios, where weather extremes are exacerbated, erroneous patterns can lead to highly biased projections. Therefore, including interpretability techniques as a diagnostic of model functioning in the evaluation process can help us to better select and design them.

1 INTRODUCTION

Global Climate Models (GCMs) are the most advanced tools available to simulate the evolution of the climate several decades into the future (climate prediction). Nonetheless, due to computational and physical limitations the outputs of these models have a typical resolution of hundreds of kilometers and do not accurately represent the regional-to-local climate variability. Regional information is required by policy makers and stakeholders to develop their adaptation plans in different future climate scenarios. To bridge this gap, several Empirical Statistical Downscaling (ESD) (Maraun & Widmann, 2018) methods have been proposed to learn an empirical relationship between a set of low-resolution variables (predictors) and the regional variable of interest (predictand). Under the *perfect-prognosis* approach (ESD-PP), the statistical models are trained following a supervised learning approach —i.e., there is a temporal synchrony between the input and output fields,— leaning on observational datasets for both the predictor (e.g reanalysis data) and the predictand (e.g., gridded observations interpolated from in-situ measurements) to infer the models.

Recently, Deep Learning (DL, (Goodfellow et al., 2016)) topologies have emerged as a promising ESD-PP technique, showing satisfactory performance to reproduce the observed local climate (Pan et al., 2019; Baño-Medina et al., 2020; Sun & Lan, 2021). Nonetheless, these models are still seen as *black boxes* in climate science, given the difficult interpretation of their inner functioning. This problem could be tackled with recent studies proposing interpretability techniques (Campos-Taberner et al., 2020; Toms et al., 2021; Dikshit & Pradhan, 2021). However the application of these techniques in the downscaling context is still incipient, with some promising results (Baño-Medina, 2020) proving that Convolutional Neural Networks (CNN, (LeCun et al., 1995)) are able to identify the relevant predictors and the spatial regions of influence thus performing some kind of automatic predictor selection in their hidden layers.

Here, we evaluate several interpretability techniques applying them to different DL models for the ESD-PP problem and show how the resulting information can be used as a diagnostic of model functioning, facilitating the interpretation of results and assisting in the identification of problems. We move beyond previous studies dealing with "medium" size domains (e.g., Europe) and test the performance of two state-of-the-art ESD-PP DL topologies of different complexities over a broad area with a strong latitudinal gradient: North America.

2 DEEP LEARNING MODELS FOR STATISTICAL DOWNSCALING

In this work we study two different CNN models with different topology previously introduced in the literature. This allows evaluating the sensitivity of the interpretability techniques for changing model complexity.

2.1 CNN10

(Baño-Medina et al., 2020) presented the first intercomparison downscaling experiment of DL topologies for ESD-PP. Following their results we adapt their best performing topology for the case of temperature (CNN10) to our study. This model is composed of three convolutional layers with 50, 25 and 10 kernels of size 3x3 and rectified linear units (ReLU) as activation function (Agarap, 2018). The output of the last convolutional layer is flattened and passed to a fully connected layer with linear activation function, whose neurons output the value for each of the gridpoints of the predictand. Despite being a simple model, authors demonstrate its overall good performance as its adaptation to out-of-domain conditions at a continental scale over Europe.

2.2 UNET-FC

In (Gadat et al., 2021) authors estimate the downscaling function of a Regional Climate Model (RCM) for the near-surface temperature of a specific domain in Western Europe. For this, they train a U-Net network (Ronneberger et al., 2015) that learns the relationship between large-scale variables and local-scale downscaled fields. This topology is composed of two different paths: the encoder and the decoder. The former takes the input and reduce its dimension, while the latter transforms the image back to desired size. A connecting path within the network keeps the encoder and decoder connected.

Based on the results of (Gadat et al., 2021) we adapt this topology to our study. The encoder of our model is formed by a succession of five blocks with a convolution, batch normalization and max pooling layer each –except the last block which does not have max pooling– with 64, 128, 256, 512 and 1024 filter maps respectively. The decoder is composed of four blocks formed by a transpose and a standard convolution –apart from the corresponding concatenation layers– of 512, 256 128 and 64 filter maps. Finally, two transpose convolutional layers and a succession of 6 convolutional layers with 64 filter maps are applied in order to get to the desired output dimensions. As activation function we use ReLU. The absence of fully connected layers made this model fully-convolutional. Notice how, unlike CNN10, this model is composed of a substantially more complex topology.

3 INTERPRETABILITY TECHNIQUES

Interpretability of DL models is still an emerging field in climate sciences. However several techniques have been tested successfully over a wide range of topologies (Simonyan et al., 2013; Zeiler & Fergus, 2014; Ribeiro et al., 2016). We select and apply three different techniques: layer-wise relevance propagation (LRP) (Bach et al., 2015), SmoothGrad (Smilkov et al., 2017) and guided backpropagation (Springenberg et al., 2014). We use the open-source library *innvestigate* (Alber et al., 2019) to implement them.

These methods assess the importance of the features of a model –pixels in the case of images– using saliency maps. These representations assign to each feature a relevance score representing its influence on the computed prediction. Different techniques use distinct approaches to compute these saliency maps. LRP recursively traces backwards the output of a neural network onto the space of

the input, identifying relevant patterns. Differently, SmoothGrad and guided backpropagation rely on the computation of the gradient of the output with respect to the input.

4 REGION OF STUDY AND DATA

Here we follow the experimental downscaling framework introduced in previous studies (Baño-Medina et al., 2020). In particular, we select as predictors 5 large-scale variables (geopotential height, zonal and meridional wind, air temperature, and specific humidity) at 4 different vertical levels (1000, 850, 700, and 500 hPa, ranging from altitudes from near surface to 5000 meters) from the ERA-Interim reanalysis (Dee et al., 2011) (trimmed to a horizontal resolution of 2°), over the region of North America (12° to 70° in latitude and -165° to -60° in longitude). This set of predictors characterizes the atmospheric configuration corresponding to a particular day. For the target predictand, we lean on the daily land near-surface air temperature over the regular gridded 0.5° EWEMBI dataset (Lange, 2019) over North America. Both datasets provide daily information for the period 1980-2008.

ERA-Interim data can be downloaded from the ECMWF website [1] and EWEMBI is available at ISIMIP [2].

5 EXPERIMENTS

The models are trained using ERA-Interim and EWEMBI datasets as low and high-resolution (input/output) pairs. The predictors variables are standardized, stacked as channels and passed to the input layer of the DL models. Models are fitted in a training set covering the period 1980-2002 by minimizing the Mean Squared Error (MSE). To avoid overfitting we follow an early stopping strategy on a random 10% split of the training set. The evaluation of these models is performed on a test set spanning the period 2003-2008.

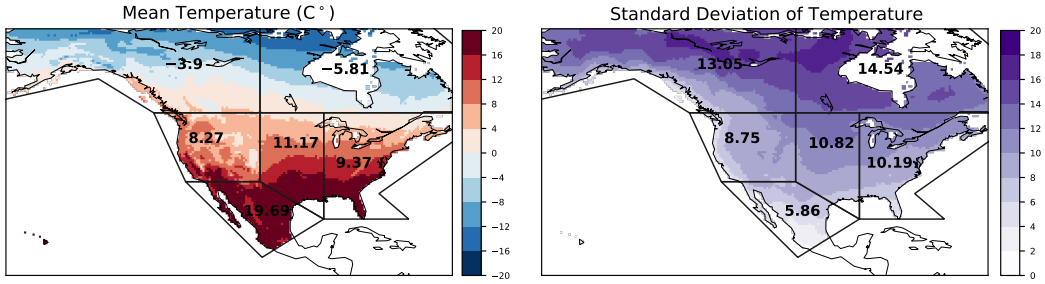

Figure 1: Mean (left) and standard deviation (right) of daily near-surface air temperature for the train period (1980-2002). Values for each of the Intergovernmental Panel on Climate Change (IPCC) reference regions (Iturbide et al., 2020) are also shown.

Note that both models try to learn the downscaling task over the full region of North America. This continent extends from the Arctic Circle to the tropics, thus suffering from an extreme latitudinal climate gradient. Figure 1 shows the mean and standard deviation of the temperature (predictand) for the train period. Mean temperature covers a wide range –from almost -20°C in the north to 20°C in the south–, and the standard deviation goes from values near 5°C in the south to almost 15°C in the north. To avoid potential spatial inhomogeneity caused by this variability, we develop a variant of CNN10 called CNN10-Stand in which the predictand is also standardized. This will allow us testing the effects of balancing the variability of the temperature across northern and southern regions.

Figure 2 shows the spatial distribution of the test RMSE for the different models over North America, as well as its seasonal evolution regarding the different IPCC reference regions. Note that the standardization required for CNN10-Stand is reversed when calculating the predictions, taking

[1] https://www.ecmwf.int/
[2] https://www.isimip.org/

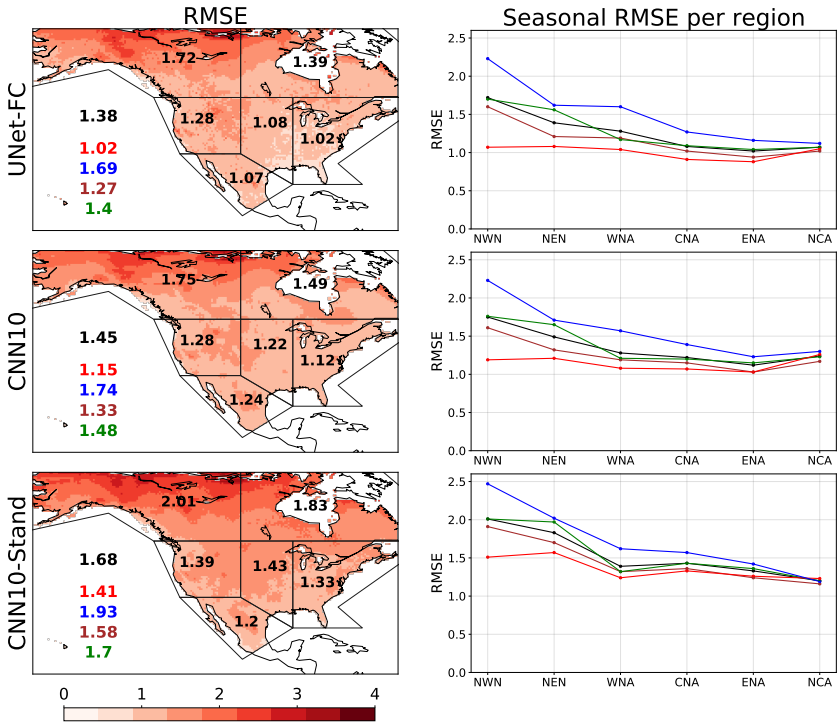

Figure 2: (left) Spatial distribution of the test RMSE (°C) for the different models (in rows): UNet-FC CNN10 and CNN10-Stand. (right) Mean seasonal test RMSE annual, summer, winter, autumn and spring values –in black, red, blue, brown and green, respectively– for the different IPCC regions (sorted from north to south and from west to east).

as reference the mean and standard deviation of the training set. UNet-FC is the best performing model, closely followed by CNN10, while CNN10-Stand shows the worst RSME values. It can be seen that for all the models the RMSE is higher in the northern regions, decreasing southwards. The RMSE across seasons is similar for all models, with the best results obtained in summer and the worst in winter.

In order to understand the role of the different spatial predictors and gain interpretability of the results, we computed the saliency maps for these models applying the three techniques previously mentioned. These are computed for each specific downscaled gridpoint (a certain output neuron) and normalized so the most influential value (a particular predictor for a particular gridpoint) takes value 1.

We have computed the mean of saliency maps for different samples –winter and summer days, full test set, month across years, etc.– observing no significant variations of the results. For CNN10 and CNN10-Stand, the three techniques (LRP, SmoothGrad and guided backpropagation) produce similar results. However, LRP fails to produce interpretable saliency maps for UNet-FC, while SmoothGrad and guided backpropagation reach similar results. This could be due to the greater complexity of the UNet-FC compared to the CNN10 models, which hampers the backward propagation of the output signal required for LRP. Due to the nature of SmoothGrad –mean gradient across variations of the same sample– it takes longer to compute than guided backpropagation. Therefore, the results for this latter technique are illustrated in the following.

Air temperature resulted the most influential predictor (in agreement with previous studies (Baño-Medina, 2020)) followed by specific humidity. Figure 3 shows the saliency maps for two different gridpoints (located in North and South extremes of the domain) for all the models. These are calculated with the samples corresponding to December across the years spanned in the test set.

For UNet-FC, the saliency maps show that, as expected, the models rely on a local region over the nearest input gridpoint. Humidity is particularly relevant in the North, but not in the South (where

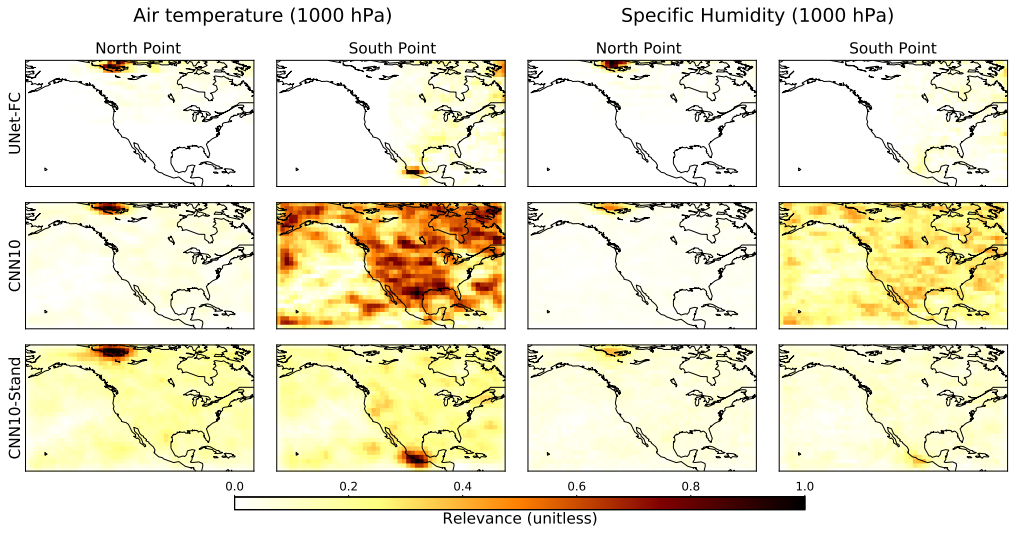

Figure 3: Mean normalized saliency maps for December observations of the test set (2003-2008) with guided backpropagation. For each model (rows) the saliency map of air temperature and specific humidity at 1000 hPa is shown for two different points (north and south), in columns.

it has smaller variability). These results provides a physical basis for the model functioning, thus enhances its credibility.

CNN10 achieves similar results to UNet-FC in terms of RMSE in both south regions. However, for the CNN10 model the saliency maps reveal a non-local pattern for the southern point, which can not be framed within a physics-based interpretation and thus indicates a potential problem with this method. This could be the effect of the smaller variance of the predictand over this regions. This is confirmed analyzing the results of CNN10-Stand (where the outputs are standardized and thus, all regions have the same weight in the error function). The saliency patterns become local, but the performance of the model is deteriorated due to the scaling factor which scales both the signal and the noise (smaller variability in tropical regions).

6 Conclusions

In this work we have analyzed different DL topologies for ESD-PP in the region of North America. These topologies, previously introduced and validated in the literature, represent different levels of complexity. Both models were previously introduced over Europe, CNN10 being applied at a continental scale. Three different interpretability techniques have been applied, all of them drawing the same conclusions about their underlying behaviour.

The saliency maps show how UNet-FC, the most complex model, learns a predictors-predictand link with a significant local pattern, a characteristic aspect of the downscaling task. CNN10, the simple model, learns a spurious relationship for the south regions which does not fit within a physics-based framework. However, the failure of CNN10 to learn this link was not reflected in the RMSE over the test period. This failure is caused by the extreme latitudinal climate gradient, which can not be learnt by the CNN10 model. By standardizing the predictand we see that the CNN10 manages to learn a plausible relationship; however, it is still unable to correctly adapt to the extreme weather of North America, as its RMSE shows. Taking as reference the RMSE, the CNN10 model would be valid for downscaling; however, when studying it also from an interpretability point of view, we see how it does not learn the authentic predictors-predictand relationship behind the downscaling task.

In climate change scenarios, where weather extremes are exacerbated, erroneous patterns can lead to highly biased projections. Including interpretability techniques in the evaluation process can help us to better select and design them, especially when applied to large continental regions with extreme climate gradients.

ACKNOWLEDGMENTS

The authors acknowledge support from Universidad de Cantabria and Consejería de Universidades, Igualdad, Cultura y Deporte del Gobierno de Cantabria via the "instrumentación y ciencia de datos para sondear la naturaleza del universo" project. J. González-Abad would also like to acknowledge the support of the funding from the Spanish *Agencia Estatal de Investigación* through the *Unidad de Excelencia María de Maeztu* with reference MDM-2017-0765.

REFERENCES

Abien Fred Agarap. Deep learning using rectified linear units (relu). *arXiv preprint arXiv:1803.08375*, 2018.

Maximilian Alber, Sebastian Lapuschkin, Philipp Seegerer, Miriam Hägele, Kristof T Schütt, Grégoire Montavon, Wojciech Samek, Klaus-Robert Müller, Sven Dähne, and Pieter-Jan Kindermans. innvestigate neural networks! *J. Mach. Learn. Res.*, 20(93):1–8, 2019.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140, 2015.

Jorge Baño-Medina. Understanding deep learning decisions in statistical downscaling models. In *Proceedings of the 10th International Conference on Climate Informatics*, pp. 79–85, 2020.

Jorge Baño-Medina, Rodrigo Manzanas, and José Manuel Gutiérrez. Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, 13(4):2109–2124, 2020.

Manuel Campos-Taberner, Francisco Javier García-Haro, Beatriz Martínez, Emma Izquierdo-Verdiguier, Clement Atzberger, Gustau Camps-Valls, and María Amparo Gilabert. Understanding deep learning in land use classification based on sentinel-2 time series. *Scientific reports*, 10(1): 1–12, 2020.

Dick P Dee, S M Uppala, Adrian J Simmons, Paul Berrisford, Paul Poli, Shinya Kobayashi, U Andrae, MA Balmaseda, G Balsamo, d P Bauer, et al. The era-interim reanalysis: Configuration and performance of the data assimilation system. *Quarterly Journal of the royal meteorological society*, 137(656):553–597, 2011.

Abhirup Dikshit and Biswajeet Pradhan. Interpretable and explainable ai (xai) model for spatial drought prediction. *Science of The Total Environment*, 801:149797, 2021.

Sébastien Gadat, Lola Corre, Antoine Doury, Aurélien Ribes, and Samuel Somot. Regional climate model emulator based on deep learning: concept and first evaluation of a novel hybrid downscaling approach. 2021.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

Maialen Iturbide, José M Gutiérrez, Lincoln M Alves, Joaquín Bedia, Ruth Cerezo-Mota, Ezequiel Cimadevilla, Antonio S Cofiño, Alejandro Di Luca, Sergio Henrique Faria, Irina V Gorodetskaya, et al. An update of ipcc climate reference regions for subcontinental analysis of climate model data: definition and aggregated datasets. *Earth System Science Data*, 12(4):2959–2970, 2020.

Stefan Lange. EartH2Observe, WFDEI and ERA-Interim data Merged and Bias-corrected for ISIMIP (EWEMBI), 2019. URL `https://dataservices.gfz-potsdam.de/pik/showshort.php?id=escidoc:3928916`. Artwork Size: 1 Files Medium: application/octet-stream Pages: 1 Files Version Number: 1.1 type: dataset.

Yann LeCun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

Douglas Maraun and Martin Widmann. *Statistical Downscaling and Bias Correction for Climate Research*. Cambridge University Press, January 2018. ISBN 978-1-108-34064-9.

Baoxiang Pan, Kuolin Hsu, Amir AghaKouchak, and Soroosh Sorooshian. Improving precipitation estimation using convolutional neural network. *Water Resources Research*, 55(3):2301–2321, 2019.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014.

Lei Sun and Yufeng Lan. Statistical downscaling of daily temperature and precipitation over china using deep learning neural models: Localization and comparison with other methods. *International Journal of Climatology*, 41(2):1128–1147, 2021.

Benjamin A Toms, Elizabeth A Barnes, and James W Hurrell. Assessing decadal predictability in an earth-system model using explainable neural networks. *Geophysical Research Letters*, 48(12): e2021GL093842, 2021.

Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pp. 818–833. Springer, 2014.