Received: 27 July 2021 | Revised: 7 February 2022 | Accepted: 8 February 2022

DOI: 10.1002/asl.1087

RESEARCH ARTICLE

Atmospheric Science Letters RMetS

A critical view on the suitability of machine learning techniques to downscale climate change projections: Illustration for temperature with a toy experiment

Alfonso Hernanz[1] | Juan Andrés García-Valero[2] | Marta Domínguez[1] | Ernesto Rodríguez-Camino[1]

[1]Spanish Meteorological Agency (AEMET), Madrid, Spain

[2]AEMET, Murcia, Spain

Correspondence
Alfonso Hernanz, Spanish Meteorological Agency (AEMET), Madrid 28040, Spain.
Email: ahernanzl@aemet.es

Abstract

Machine learning is a growing field of research with many applications. It provides a series of techniques able to solve complex nonlinear problems, and that has promoted their application for statistical downscaling. Intercomparison exercises with other classical methods have so far shown promising results. Nevertheless, many evaluation studies of statistical downscaling methods neglect the analysis of their extrapolation capability. In this study, we aim to make a wakeup call to the community about the potential risks of using machine learning for statistical downscaling of climate change projections. We present a set of three toy experiments, applying three commonly used machine learning algorithms, two different implementations of artificial neural networks and a support vector machine, to downscale daily maximum temperature, and comparing them with the classical multiple linear regression. We have tested the four methods in and out of their calibration range, and have found how the three machine learning techniques can perform poorly under extrapolation. Additionally, we have analysed the impact of this extrapolation issue depending on the degree of overlapping between the training and testing datasets, and we have found very different sensitivities for each method and specific implementation.

KEYWORDS

climate projections, evaluation, extrapolation, machine learning, neural networks, statistical downscaling, support vectors

1 | INTRODUCTION

Global climate models (GCMs) are the main tool to simulate future climate projections, but their coarse resolution makes them unsuitable for the local scale (Charles et al., 2004; Schoof, 2013; Wilby et al., 2004). This gap is often filled by applying some type of downscaling over GCMs outputs (Maraun et al., 2010). The two primary categories of downscaling techniques are (1) dynamic downscaling, mostly by nesting high resolution regional climate models (RCMs) in GCMs and (2) statistical downscaling (SD), by establishing empirical/statistical relationships between large-scale predictors and local surface predictands. SD relies on the following assumptions: (1) stationarity of these relationships under climate change (Wilby et al., 2004); (2) predictors are reliably simulated by GCMs; (3) predictors contain climate change signal; and (4) predictors are strongly correlated with predictands (Wilby et al., 2004). There are many different approaches to SD, each with its strengths and weaknesses, and the VALUE EU COST Action (Maraun et al., 2015) established a framework to evaluate and intercompare them.

Machine learning (ML) techniques are able to simulate complex nonlinear relationships, what has promoted their application as statistical downscaling models (SDMs) (see, e.g., Goyal et al., 2012; Li et al., 2020; Sachindra et al., 2018; Vandal et al., 2019). Nevertheless, Hsieh (2009) pointed out the risk of using ML techniques to downscale climate change projections. While they have proved able to map complex functions, they can exhibit significantly different behaviours in and out of their calibration range.

Many evaluation studies limit to the so-called *Perfect Predictor* experiment (Maraun & Widmann, 2018; Maraun et al., 2015), which usually does not explicitly test the behaviour of SDMs under extrapolation, and never does it under such a degree of extrapolation as it is expected in the future. The ability of SDMs to extrapolate can be evaluated in different ways. A first option consists in splitting a historical dataset in training/testing samples in a specific way, for example by leaving the warmest/driest years for validation. This analysis allows to explore a certain degree of extrapolation, but usually narrower than the one expected in future projections. Some examples of this approach can be seen in Gutiérrez et al. (2013), San-Martín et al. (2017) and Baño-Medina et al. (2020). Another approach consists of using of RCMs as pseudo-observations to train SDMs in a historical period and evaluate them in the future, the so-called *pseudo-reality* experiments (Maraun & Widmann, 2018; Maraun et al., 2015). These experiments check the necessary (but not sufficient) condition of SDMs being able to reproduce climate change signal given by GCM-RCMs, and must be limited to those variables which are realistically simulated by GCM-RCMs. Although this approach allows the analysis of a wider range of extrapolation, it introduces a source of uncertainty associated with RCMs. Some examples of this approach can be found in Charles et al. (1999), Vrac et al. (2007), Gaitan et al. (2014) and Hernanz et al. (2021). And finally, another possible approach consists in comparing raw GCMs with SDMs, which can only be done by averaging over large enough areas. This approach allows to analyse whether SDMs are able to preserve trends projected by GCMs, but it cannot tackle imperfections on the finer spatial scales. Some examples of this approach can be found in Vandal et al. (2019), Xu et al. (2020) and Baño-Medina et al. (2021). Although each approach has its drawbacks, these types of studies are very valuable and can reveal significant extrapolation issues.

In this paper, we perform a set of three toy experiments in order to show how some ML techniques commonly used to downscale climate change projections can behave extremely wrong under extrapolation. The main objective of this paper is to make a wake-up call to the community on the potential risks of using these techniques, to reinforce the need of some kind of extrapolation analysis as a key piece of SDMs evaluation studies and to analyse the sensitivity of three common ML SDMs to the degree of overlapping between the training and testing datasets. The paper is organised as follows. First, a description of the datasets used is given at Section 2, followed by a brief introduction to the downscaling methods in Section 3. The experiment design is explained in Section 4. And finally, results and concluding remarks are shown and discussed in Sections 5 and 6, respectively.

2 | DATA

In order to keep this study as simple and clear as possible, it has been limited to the downscaling of daily maximum surface temperature (TMAX) using temperature at 850 hPa (T850) as the only predictor. This predictor meets three out of the four conditions enumerated in

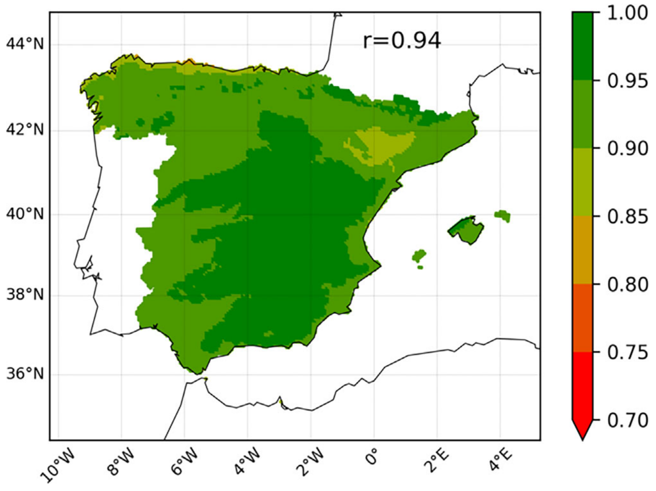

FIGURE 1 Spatial distribution of Pearson correlation coefficient between T850 (interpolated from the reanalysis ERA-interim) and daily maximum temperature (from the high resolution AEMET grid [0.05°]). See text for details

Section 1: reliability by GCMs, signal of climate change and strong correlation with predictand (see Figure 1). Note that this simple choice of T850 as the unique predictor is only reasonable when near surface and free atmosphere are strongly coupled.

T850 has been taken from the reanalysis ERA-Interim (Dee et al., 2011) of the European Centre for Medium-Range Weather Forecasts (ECMWF) for the domain (45° N, 34.5° N, 10.5° W, 4.5° E) with spatial resolution of 1.5° × 1.5° and as daily mean values (from 00, 06, 12 and 18 UTC). It has been interpolated to each target point with a bilinear interpolation of the four nearest neighbours and standardised so it is used in the form of anomalies.

TMAX has been taken from a high resolution grid (0.05°) consisting of 16,156 points over Spain (mainland and Balearic Islands) developed by AEMET (Peral et al., 2017). This grid has been generated using an adaptation of the HIRLAM Surface Analysis code (Navascués et al., 2003; Rodríguez et al., 2003), based on an Optimum Interpolation algorithm (Daley, 1991), applied over a selection of 1800 stations from the AEMET observational network.

Both datasets cover the period 1980–2016, and the training/testing split has been performed in different ways for each of the three experiments explained in Section 4.

3 | DOWNSCALING METHODS

For these experiments we have applied three commonly used ML techniques, two different implementations of artificial neural network (ANN) and a support vector regression (SVR), and we have compared their results with the classical method of multiple linear regression (MLR). Note that, since this particular problem is one-dimensional, MLR really corresponds to a simple linear regression. For their implementation, we have used the *Python* machine learning library *Scikit-learn* (Pedregosa et al., 2011).

ANNs (McCulloch & Pitts, 1943; Rosenblatt, 1958) are supervised learning algorithms based on the biological neurons' behaviour, imitating them by nodes which work as *perceptrons* (Rosenblatt, 1958). In the popular implementation of ANNs called *multilayer perceptron* (MLP), these nodes are organised in several layers, the input layer, the output layer and a set of hidden layers, which communicate with adjacent layers. Each node receives several input signals (x_j) from the (m) nodes at the previous layer and adds them with different weights (w_j) (Equation 1). This resulting input signal (z) is then fed to an activation function (g), so the node will pass a signal to the next layer depending on whether the activation function exceeds a certain threshold or not. For these experiments, we have used two different implementations of ANN, one using a *rectified linear unit activation function* (ANN-RELU) and the other using a logistic activation function (ANN-LOG)

$$z = \sum_{j=1}^{m} w_j x_j. \tag{1}$$

The training of an MLP consists in searching those weights which minimise errors, and it is usually achieved by an iterative process in which signals are transmitted forward, errors are propagated backwards and weights are updated until a certain condition is fulfilled. During this iterative process, ANNs algorithms can get trapped in local minima, which is one of their major drawbacks, along with their high computational training cost. ANNs can tackle both classification and regression problems, and they have been extensively applied to SD (Chadwick et al., 2011; Coulibaly et al., 2005; Dibike & Coulibaly, 2006; Mendes & Marengo, 2013; Sailor et al., 2000; Snell et al., 2000; Trigo & Palutikof, 1999).

SVMs (Boser et al., 1992; Cortes & Vapnik, 1995; Vapnik, 1995) are supervised learning algorithms originally designed for classification based on a combination of minimal errors and maximising distances from data points to the boundary decision (maximal margins). A slightly different version of the algorithm, SVR, can be applied to regression problems. Linear SVR (Drucker et al., 1997) searches for the optimum hyperplane (defined by its parameters w, w_0) by penalysing only errors greater than a certain threshold (ε), which defines the so-called *ε-tube*. The nonlinear problem is tackled by mapping original data (x), through a transformation (φ), to a higher dimensional space (*feature space*) where the problem becomes linear. Thus, an SVR corresponds to the hyperplane in the *feature space*:

$$y = w^T \varphi(x) + w_0. \tag{2}$$

And the combination of minimal errors and maximal margins corresponds to minimising

$$\frac{1}{2} w^T w + C \sum_{i=1}^{n} (\xi_i + \xi_i{}^*), \tag{3}$$

subject to

$$\begin{gathered} y_i - \left(w^T \varphi(x_i) + w_0\right) \leq \varepsilon + \xi_i, i = 1, 2, \ldots, n, \\ \left(w^T \varphi(x_i) + w_0\right) - y_i \leq \varepsilon + \xi_i{}^*, i = 1, 2, \ldots, n, \\ \xi_i, \xi_i{}^* \geq 0, i = 1, 2, \ldots, n, \end{gathered} \tag{4}$$

where ξ_i and ξ_i^*, called slack variables, are the penalties corresponding to errors out of the *ε-tube*, x_i, y_i correspond to pairs of data points and C is a hyperparameter that establishes the balance between errors and maximal margin. For a more detailed description on SVR, see Drucker et al. (1997). Expensive high-dimensional computations, usually referred to as *the curse of dimensionality*, are avoided by the so-called *kernel trick* (see Shawe-Taylor & Cristianini, 2004), which consists in visiting the *feature space* exclusively to compute *inner products* of pairs of points without explicitly mapping each point individually. This is only possible by carefully defining the inner product as a specific function called *kernel (K)* with some desired properties. For this work, we have used a *radial basis function kernel* of variance σ^2

$$K(x_i,x_j)=\varphi(x_i)^T\varphi(x_j)=e^{-\frac{\|x_i-x_j\|^2}{\sigma^2}}. \qquad (5)$$

Finally, the problem consists of solving a convex quadratic programming problem, or a set of linear equations in the *least-square support vector machine* variant (Suykens & Vandewalle, 1999), both of them lacking the inconvenience of local minimum. Different forms of SVMs have been widely applied to SD, both for classification and regression (Anandhi et al., 2008; Chen et al., 2010; Ghosh & Mujumdar, 2008; Hou et al., 2014; Sachindra et al., 2013; Tripathi et al., 2006; Yu & Liong, 2007).

4 | EXPERIMENT DESIGN

In order to prove how ML techniques can exhibit strange behaviours out of their calibration range and to analyse their sensitivity to the degree of extrapolation, three toy experiments have been performed (Figure 2).

Experiment 1: “full overlapping”—for each grid point, the training/testing split is performed randomly in a 60/40 ratio. The aim of this first experiment is to evaluate and intercompare the three methods when no extrapolation takes place.

Experiment 2: “no overlapping”—for each grid point, the 60th percentile of T850 is used as a threshold for the training/testing split, using values under it for training and values over it for testing. The aim of this experiment is to compare the three methods under extrapolation, in the extreme and non-realistic case of zero overlapping between the training and testing datasets.

Experiment 3: “partial overlapping”—for each grid point, the training/testing split is performed randomly in a 60/40 ratio as in Experiment 1. But the testing dataset is then modified by shifting predictors to higher values

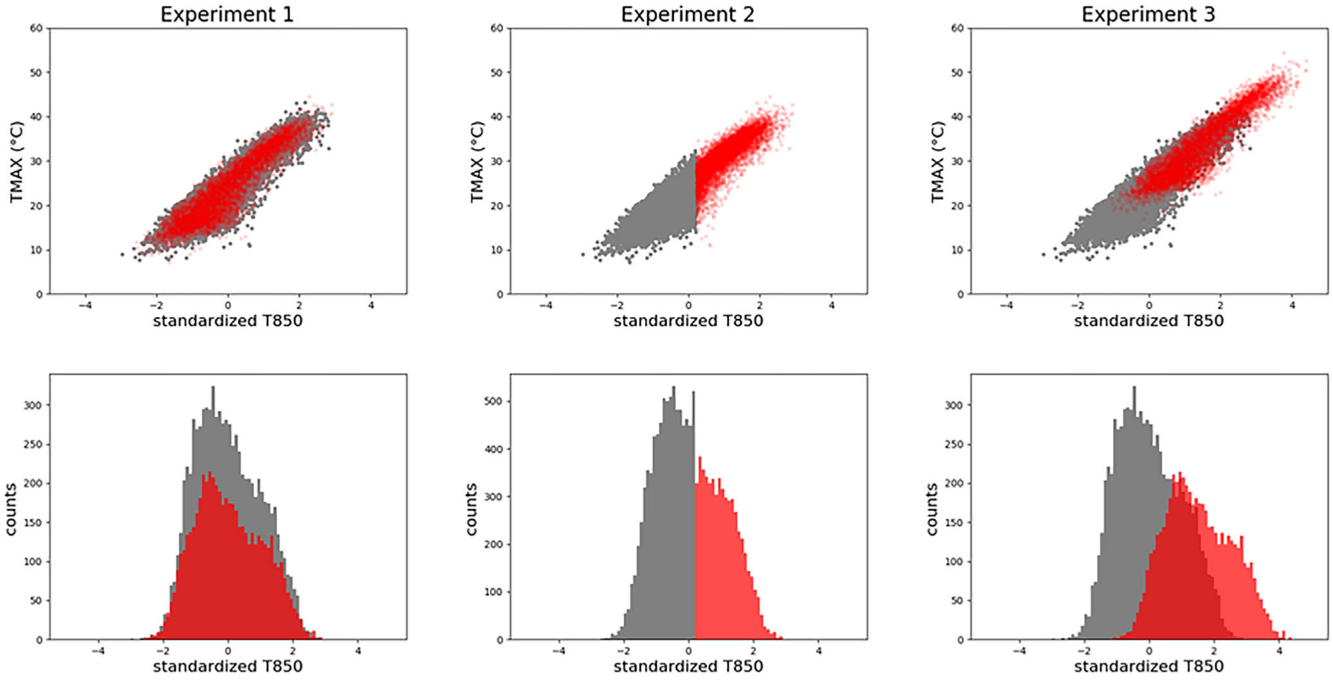

FIGURE 2 Illustrative scheme of the training (grey) and testing (red) datasets, for a single grid point. Experiment 1: “full overlapping” (first column), Experiment 2: “no overlapping” (second column) and Experiment 3: “partial overlapping” (third column). Of the different shifts applied in Experiment 3 (ranging from +0.5 to +4 standard deviations), this figure corresponds to the specific shift of +1.5 standard deviations. The upper row shows standardised T850 (x-axis) versus TMAX (°C, y-axis), and the lower row shows the frequency (counts) of T850

and using the estimates from MLR as predictands, since in experiment 3 there are no observed predictands. Different shifts have been applied: +0.5, +1, +1.5, …, +4 standard deviations. This experiment aims to analyse the sensitivity of each method to the level of extrapolation, with a more realistic approach than Experiment 2 in terms of shape of the training and testing distributions and degree of overlapping between them. On the other hand, it is important to remark that this experiment relies on the assumption of linear relationship between T850 and TMAX, for it validates against linearly extrapolated data instead of against real data. But given the high linear correlation between T850 and TMAX in the observed range (Figure 1) it seems a reasonable assumption to be made for the purpose of this experiment.

In the three experiments, 8110 days are used for training and 5405 for testing, and the hyperparameter tuning is performed by cross-validation using exclusively the training data set. SDMs are validated with the testing dataset by their root mean square error (RMSE).

Additionally, for Experiment 2, SDMs have been applied over both the training and testing datasets, in order to visualise their behaviour out of the calibration range.

5 | RESULTS

In the first experiment, all SDMs achieve similar RMSEs, with slightly better results for ML techniques (Figure 3). Nonlinear methods do not add much value here because of the high linear correlation among predictor and predictand (Figure 1). Nevertheless, these very same methods have proved to overcome MLR when using a greater set of predictors (Hernanz et al., 2021).

When applied under extreme extrapolation, ML techniques display very high RMSEs compared to MLR (Figure 3). Despite the similar RMSEs of the four SDMs in Experiment 1, their RMSEs in Experiment 2 differ significantly. Whereas the mean RMSE goes from 2.58°C in Experiment 1 to 3.29°C in Experiment 2 for MLR, for ANA-RELU it goes from 2.51 to 6.06°C, for ANA-LOG it goes from 2.49 to 7.66°C and for SVR it goes from 2.51 to 5.52°C. Some regions present much larger RMSEs than in Experiment 1, while in other regions RMSEs barely increase. A possible explanation comes with the fact that regions with larger RMSEs correspond to some of the main Spanish valleys with high occurrence of fog, especially in winter, which strongly condition daily maximum temperatures and could lead to a certain uncoupling between T850 and TMAX. A further analysis of this evidence is out of the scope of this work, but it might constitute an interesting future study.

An illustrative example of the different behaviours by the four methods under extrapolation, over a single grid point, is shown in Figure 4. The four methods map a similar relationship between T850 and TMAX over the training sample, but their behaviours diverge largely under extrapolation, where ANN-RELU, ANN-LOG and SVR do not perform well.

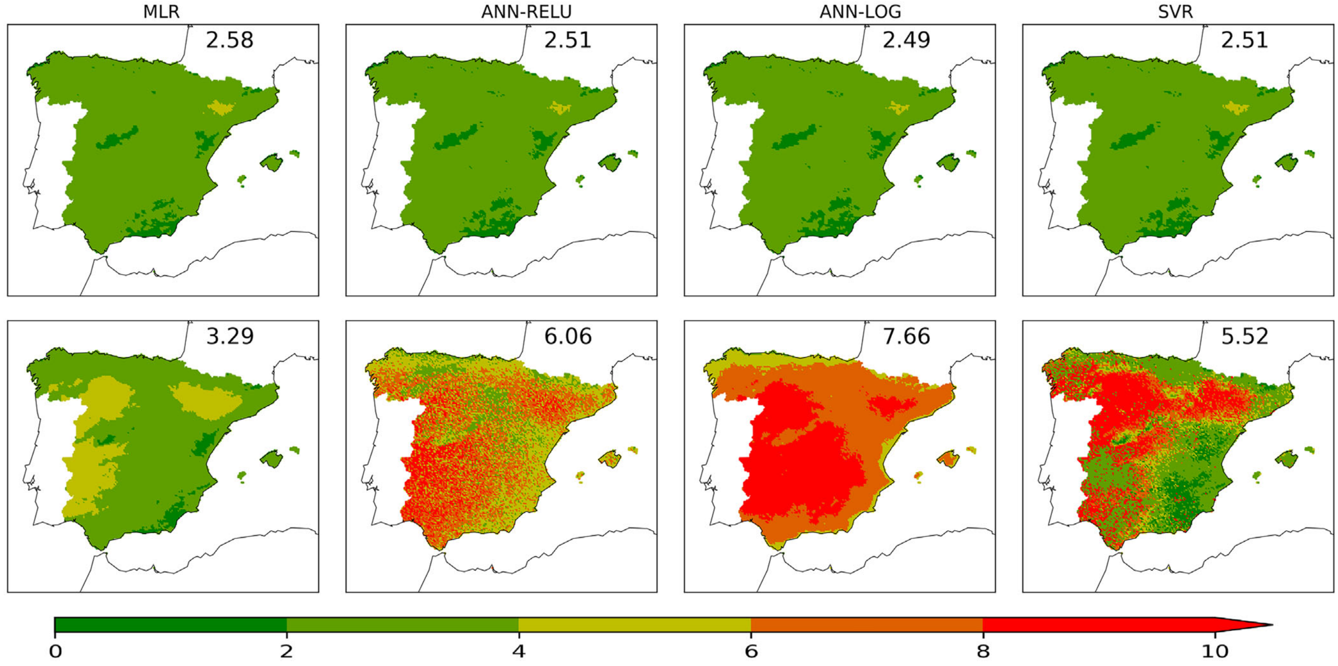

FIGURE 3 Spatial distribution and spatial mean of RMSE (°C) for daily maximum temperature in Experiments 1 (upper row) and 2 (lower row) by MLR (first column), ANN-RELU (second column), ANN-LOG (third column) and SVR (fourth column)

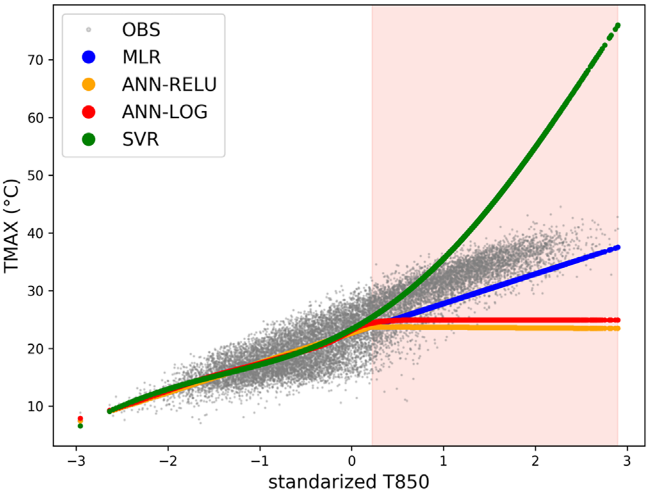

FIGURE 4 Standardised T850 versus observed (grey) and downscaled maximum temperature (°C) by MLR (blue), ANN-RELU (orange), ANN-LOG (red) and SVR (green) over the training (white background) and testing (red background) datasets for Experiment 2. This figure represents an illustrative example over a single grid point of coordinates 1.730° W and 37.318° N

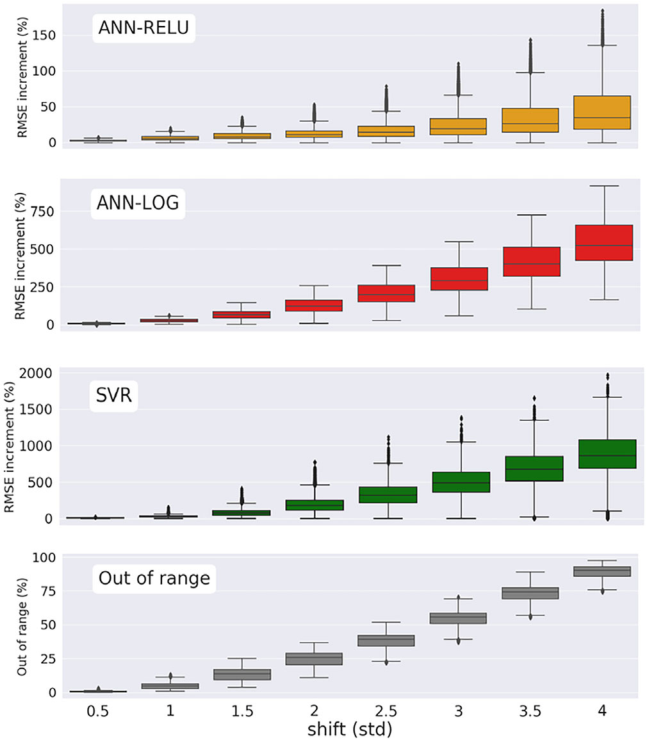

FIGURE 5 Increment of RMSE (%) for ANN-RELU (first panel, orange), ANN-LOG (second panel, red) and SVR (third panel, green) at Experiment 3, for different shifts applied to the testing dataset. The fourth panel (grey) shows, for each shift, the amount of testing data (%) that lies out of the calibration range. Each box contains the quartiles of all grid points (16,156 data) and the whiskers extend to a maximum of 1.5 times the interquartile range. Outliers beyond this range are plotted individually. Note that vertical scales are different for each panel

Finally, Figure 5 shows results from Experiment 3, in which the three ML methods are evaluated for different degrees of extrapolation. MLR is not evaluated in this experiment, because the linear relationship has been used to build the shifted testing datasets. While the three methods displayed extrapolation problems of the same order in Experiment 2 (Figure 3), ANN-RELU behaves much better than ANN-LOG and SVR under extrapolation in this third experiment, in which the training and testing distributions overlap in a more realistic way. For example, for a shift of +0.5 standard deviations and a percentage of testing data out of the calibration range of 0%–3%, RMSEs by ANN-RELU increment up to 6% compared to Experiment 1, but for ANN-LOG and SVR these increments go up to around 20%. For a shift of +1 standard deviation (0%–13% of data out of range), ANN-RELU increments its RMSEs up to 19%, while ANN-LOG and SVR reach 63% and 147%, respectively. And for a shift of +2 standard deviations (10%–36% of data out of range), ANN-RELU goes up to 53% increments, while ANN-LOG and SVR reaches increments of 263% and 775%.

6 | CONCLUDING REMARKS

ML is a growing field with many applications in atmospheric sciences, being SD of climate change projections one of them. In this study, we have analysed the behaviour of three commonly used ML techniques, two different implementations of ANN (ANN-RELU and ANN-LOG) and an SVR, under extrapolation, through a set of three toy experiments: the first one to evaluate them when no extrapolation takes place, the second one to evaluate them in the extreme case of no overlapping between the training and testing datasets and the third one to analyse their sensitivity to the degree of overlapping. We have proved how the three ML SDMs can behave extremely wrong out of their calibration range, despite their scores inside the calibration range being as good as or even better than those for MLR. Then, we have analysed the impact of this potential extrapolation issue depending on the degree of overlapping between the training and testing datasets. We have found that ANN-RELU errors, when some degree of overlapping takes place, are much lower than those of ANN-LOG and SVR, which has revealed to be extremely sensitive to extrapolation.

This set of experiments has allowed us to prove how three commonly used ML techniques can perform wrong under extrapolation, so their suitability for SD of climate change projections should be seriously questioned. Furthermore, we have seen how, for a specific technique (ANN), extrapolation issues are very sensitive to the particular implementation. For this technique, we have compared two commonly used activation functions, but other architectures and implementations are possible and also common. For this reason, results shown here might not be straightforward generalised to other techniques or specific implementations, and the extrapolation capability of these methods should be thoroughly examined for each case (variable, region, set of predictors, architecture, etc.).

It is important to clarify that these experiments do not intend to replace other more realistic evaluation approaches which also tackle the extrapolation issue, like the ones mentioned in Section 1. Nonetheless, it is worth noting to remark that experiments which validate over spatially/temporarily aggregated data might hide extrapolation problems in finer spatial/temporal scales. Thus, we consider that analysing the response of ML SDMs through synthetic extrapolation experiments like these ones constitutes a good practice and a first recommended step to detect and be aware of their potential risks.

Finally, we would like to point to the emerging field of *Physics-Constrained Machine Learning* (see, e.g., Willard et al., 2020; Kashinath et al., 2021) as a possible way of alleviating these extrapolation problems, by reducing the high amount of degrees of freedom that these techniques cope with.

ACKNOWLEDGEMENTS

The authors would like to acknowledge all the team involved in the ERA-Interim reanalysis and in the development of the Python machine learning library Scikit-learn. They would like to thank María Candelas Peral García and Beatriz Navascués for their development of the high resolution observational grid over Spain here used, and to the numerous altruistic collaborators of the AEMET observational network. They would also like to thank two anonymous reviewers for their suggestions, which have improved the quality of the document.

AUTHOR CONTRIBUTIONS

Alfonso Hernanz: Software; visualization; writing – original draft. **Juan Andrés García-Valero:** Writing – review and editing. **Marta Domínguez:** Writing – review and editing. **Ernesto Rodríguez-Camino:** Supervision; writing – review and editing.

ORCID

Alfonso Hernanz https://orcid.org/0000-0003-1091-0422
Juan Andrés García-Valero https://orcid.org/0000-0002-3914-6328
Marta Dominguez https://orcid.org/0000-0001-7840-5516
Ernesto Rodríguez-Camino https://orcid.org/0000-0002-1565-2373

REFERENCES

Anandhi, A., Srinivas, V.V., Nanjundiah, R.S. & Nagesh Kumar, D. (2008) Downscaling precipitation to river basin in India for IPCC SRES scenarios using support vector machine. *International Journal of Climatology*, 28, 401–420. https://doi.org/10.1002/joc.1529

Baño-Medina, J., Manzanas, R. & Gutiérrez, J.M. (2020) Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, 13, 2109–2124. https://doi.org/10.5194/gmd-13-2109-2020

Baño-Medina, J., Manzanas, R. & Gutiérrez, J.M. (2021) On the suitability of deep convolutional neural networks for continental-wide downscaling of climate change projections. *Climate Dynamics*, 57, 2941–2951. https://doi.org/10.1007/s00382-021-05847-0

Boser, B., Guyon, I. and Vapnik, V. (1992). A training algorithm for optimal margin classifier. Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory. 144–152. https://doi.org/10.1145/130385.130401

Chadwick, R., Coppola, E. & Giorgi, F. (2011) An artificial neural network technique for downscaling GCM outputs to RCM spatial scale. *Nonlinear Processes in Geophysics*, 18, 1013–1028. https://doi.org/10.5194/npg-18-1013-2011

Charles, S., Bates, B., Whetton, P. & Hughes, J. (1999) Validation of downscaling models for changed climate conditions: case study of southwestern Australia. *Climate Research*, 12, 1–14. https://doi.org/10.3354/cr012001

Charles, S.P., Bates, B.C., Smith, I.N. & Hughes, J.P. (2004) Statistical downscaling of daily precipitation from observed and modelled atmospheric fields. *Hydrological Processes*, 18, 1373–1394. https://doi.org/10.1002/hyp.1418

Chen, H., Guo, J., Xiong, W., Guo, S. & Xu, C.-Y. (2010) Downscaling GCMs using the smooth support vector machine method to predict daily precipitation in the Hanjiang Basin. *Advances in Atmospheric Sciences*, 27, 274–284. https://doi.org/10.1007/s00376-009-8071-1

Cortes, C. & Vapnik, V. (1995) Support-vector networks. *Machine Learning*, 20, 273–297. https://doi.org/10.1007/BF00994018

Coulibaly, P., Dibike, Y.B. & Anctil, F. (2005) Downscaling precipitation and temperature with temporal neural networks. *Journal of Hydrometeorology*, 6(4), 483–496. https://journals.ametsoc.org/view/journals/hydr/6/4/jhm409_1.xml

Daley, R. (1991) *Atmospheric data analysis. Cambridge atmospheric and space science series*. Cambridge: Cambridge University Press, p. 341. https://doi.org/10.1002/joc.3370120708

Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S. et al. (2011) The ERA-interim reanalysis: configuration and performance of the data assimilation system. *Quarterly Journal of the Royal Meteorological Society*, 137, 553–597. https://doi.org/10.1002/qj.828

Dibike, Y.B. & Coulibaly, P. (2006) Temporal neural networks for downscaling climate variability and extremes. *Neural Networks*, 19(2), 135–144. https://doi.org/10.1016/j.neunet.2006.01.003

Drucker, H., Burges, C.J., Kaufman, L., Smola, A. & Vapnik, V. (1997) Support vector regression machines. *Advances in Neural Information Processing Systems*, 9, 155–161.

Gaitan, C., Hsieh, W. & Cannon, A. (2014) Comparison of statistically downscaled precipitation in terms of future climate indices and daily variability for southern Ontario and Quebec, Canada. *Clim Dyn*, 43, 1–17. https://doi.org/10.1007/s00382-014-2098-4

Ghosh, S. & Mujumdar, P. (2008) Statistical downscaling of GCM simulations to streamflow using relevance vector machine. *Advances in Water Resources*, 31, 132–146. https://doi.org/10.1016/j.advwatres.2007.07.005

Goyal, M.K., Burn, D.H. & Ojha, C.S.P. (2012) Evaluation of machine learning tools as a statistical downscaling tool: temperatures projections for multi-stations for Thames River basin, Canada. *Theoretical and Applied Climatology*, 108, 519–534. https://doi.org/10.1007/s00704-011-0546-1

Gutiérrez, J.M., San-Martín, D., Brands, S., Manzanas, R. & Herrera, S. (2013) Reassessing statistical downscaling techniques for their robust application under climate change conditions. *Journal of Climate*, 26(1), 171–188. https://doi.org/10.1175/JCLI-D-11-00687.1

Hernanz, A., García-Valero, J.A., Domínguez, M., Ramos-Calzado, P., Pastor-Saavedra, M.A. & Rodríguez-Camino, E. (2021) Evaluation of statistical downscaling methods for climate change projections over Spain: present conditions with perfect predictors. *International Journal of Climatology*, 42, 762–776. https://doi.org/10.1002/joc.7271

Hernanz, A., García-Valero, J.A., Domínguez, M. & Rodríguez-Camino, E. (2021) Evaluation of statistical downscaling methods for climate change projections over Spain: future conditions with pseudo reality (transferability experiment). *International Journal of Climatology*, 1–14. https://doi.org/10.1002/joc.7464

Hou, Y., Huang, X., Chen, H. & Xu, C.-Y. (2014) A statistical downscaling method based on least squares support vector machines. *Journal of Water Resources Research (in Chinese with English abstract)*, 3, 72–77. https://doi.org/10.12677/JWRR.2014.31012

Hsieh, W. (2009) *Machine learning methods in the environmental sciences: neural networks and kernels*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511627217

Kashinath, K., Mustafa, M., Albert, A., Wu, J.L., Jiang, C., Esmaeilzadeh, S. et al. (2021) Physics-informed machine learning: case studies for weather and climate modelling. *Philosophical Transactions of the Royal Society A*, 379(2194), 20200093.

Li, X., Li, Z., Huang, W. & Zhou, P. (2020) Performance of statistical and machine learning ensembles for daily temperature downscaling. *Theoretical and Applied Climatology*, 140, 571–588. https://doi.org/10.1007/s00704-020-03098-3

Maraun, D. & Widmann, M. (2018) *Statistical downscaling and bias correction for climate research*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781107588783

Maraun, D., Wetterhall, F., Ireson, A.M., Chandler, R.E., Kendon, E.J., Widmann, M. et al. (2010) Precipitation downscaling under climate change: recent developments to bridge the gap between dynamical models and the end user. *Reviews of Geophysics*, 48(3). https://doi.org/10.1029/2009RG000314

Maraun, D., Widmann, M., Gutiérrez, J.M., Kotlarski, S., Chandler, R.E., Hertig, E. et al. (2015) Value: a framework to validate downscaling approaches for climate change studies. *Earth's Future*, 3(1), 1–14. https://doi.org/10.1002/2014EF000259

McCulloch, W.S. & Pitts, W. (1943) A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, 5, 115–133. https://doi.org/10.1007/BF02478259

Mendes, D. & Marengo, J. (2013) Temporal downscaling: a comparison between artificial neural network and autocorrelation techniques over the Amazon Basin in present and future climate change scenarios. *Theoretical and Applied Climatology*, 100, 413–421. https://doi.org/10.1007/s00704-009-0193-y

Navascués, B., Rodríguez, E., Ayuso J.J., and Järvenoja, S (2003). Analysis of surface variables and parameterization of surface processes in HIRLAM. Part II: Seasonal assimilation experiment. HIRLAM Technical Report 59. http://hdl.handle.net/20.500.11765/12003

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O. et al. (2011) Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, 12, 2825–2830. https://doi.org/10.5555/1953048.2078195

Peral, C., Navascués, B., and Ramos, P. (2017). Serie de precipitación diaria en rejilla con fines climáticos. Nota Técnica no. 24 AEMET. http://www.aemet.es/documentos/es/conocermas/recursos_en_linea/publicaciones_y_estudios/publicaciones/NT_24_AEMET/NT_24_AEMET.pdf

Rodríguez, E., Navascués, B., Ayuso, J.J., and Järvenoja, S. (2003). Analysis of surface variables and parameterization of surface processes in HIRLAM. Part I: Approach and verification by parallel runs. HIRLAM Technical Report 58. http://hdl.handle.net/20.500.11765/12004

Rosenblatt, F. (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65, 386–408.

Sachindra, D.A., Huang, F., Barton, A. & Perera, B.J.C. (2013) Least square support vector and multi-linear regression for statistically downscaling general circulation model outputs to catchment streamflows. *International Journal of Climatology*, 33, 1087–1106. https://doi.org/10.1002/joc.3493

Sachindra, D.A., Ahmed, K., Rashid, M., Shahid, S. & Perera, B.J.C. (2018) Statistical downscaling of precipitation using machine learning techniques. *Atmospheric Research*, 212, pags 240-258. https://doi.org/10.1016/j.atmosres.2018.05.022

Sailor, D., Hu, T., Li, X. & Rosen, J.N. (2000) A neural network approach to local downscaling of GCM output for assessing wind power implications of climate change. *Renewable Energy*, 19, 359–378. https://doi.org/10.1016/S0960-1481(99)00056-7

San-Martín, D., Manzanas, R., Brands, S., Herrera, S. & Gutiérrez, J.M. (2017) Reassessing model uncertainty for regional projections of precipitation with an ensemble of statistical downscaling methods. *Journal of Climate*, 30(1), 203–223. https://doi.org/10.1175/JCLI-D-16-0366.1

Schoof, J.T. (2013) Statistical downscaling in climatology. *Geography Compass*, 7(4), 249–265. https://doi.org/10.1111/gec3.12036

Shawe-Taylor, J. & Cristianini, N. (2004) *Kernel methods for pattern analysis*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511809682

Snell, S.E., Gopal, S. & Kaufmann, R.K. (2000) Spatial interpolation of surface air temperatures using artificial neural networks: evaluating their use for downscaling GCMs. *Journal of Climate*, 13(5), 886–895.

Suykens, J. & Vandewalle, J. (1999) Least squares support vector machine classifiers. *Neural Processing Letters*, 9, 293–300. https://doi.org/10.1023/A:1018628609742

Trigo, R. & Palutikof, J. (1999) Simulation of daily temperatures for climate change scenarios over Portugal: a neural network model approach. *Climate Research*, 13, 45–59. https://doi.org/10.3354/cr013045

Tripathi, S., Venkata, S. & Nanjundiah, R. (2006) Downscaling of precipitation for climate change scenarios: a support vector machine approach. *Journal of Hydrology*, 330(3), 621–640. https://doi.org/10.1016/j.jhydrol.2006.04.030

Vandal, T., Kodra, E. & Ganguly, A.R. (2019) Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation. *Theoretical and Applied Climatology*, 137, 557–570. https://doi.org/10.1007/s00704-018-2613-3

Vapnik, V.N. (1995) *The nature of statistical learning theory*. Berlin: Springer-Verlag.

Vrac, M., Stein, M., Hayhoe, K. & Liang, X.-Z. (2007) A general method for validating statistical downscaling methods under future climate change. *Geophysicial Research Letters.*, 34, L18701. https://doi.org/10.1029/2007GL030295

Wilby, R., Charles, S., Zorita, E., Timbal, B., Whetton, P. and Mearns, L. (2004). Guidelines for use of climate scenarios developed from statistical downscaling methods. Supporting material of the Intergovernmental Panel on Climate Change.

Willard, J., Jia, X., Xu, S., Steinbach, M., and Kumar, V. (2020). Integrating physics-based modeling with machine learning: a survey. arXiv:2003.04919. https://doi.org/10.1145/1122445.1122456

Xu, R., Chen, N., Chen, Y. & Chen, Z. (2020) Downscaling and projection of multi-CMIP5 precipitation using machine learning methods in the upper Han River basin. *Advances in Meteorology*, 2020, 8680436. https://doi.org/10.1155/2020/8680436

Yu, X. & Liong, S.-Y. (2007) Forecasting of hydrologic time series with ridge regression in feature space. *Journal of Hydrology*, 332, 290–302. https://doi.org/10.1016/j.jhydrol.2006.07.003

How to cite this article: Hernanz, A., García-Valero, J. A., Domínguez, M., & Rodríguez-Camino, E. (2022). A critical view on the suitability of machine learning techniques to downscale climate change projections: Illustration for temperature with a toy experiment. *Atmospheric Science Letters*, e1087. https://doi.org/10.1002/asl.1087