Potential journals

1) Environmental Research Letters
2) Earth's future (AGU) Submit here
3) Water Resources Research
4) Atmospheric Research

Cutting the cost and time of dynamical downscaling with machine learning

Sanaa Hobeichi[1,2], Nidhi Nishant[1,2], Yawen Shao[2,3], Gab Abramowitz[1,2], Andy Pitman[1,2], Steve Sherwood[1,2], Craig Bishop[2,3], and Sam Green[1,2]

[1] Climate Change Research Centre, UNSW Sydney, NSW 2052, Australia.

[2]ARC Centre of Excellence for Climate Extremes, UNSW Sydney, NSW 2052, Australia.

[3] School of Geography, Earth and Atmospheric Sciences, The University of Melbourne, Parkville, Victoria, Australia

Correspondence: Sanaa Hobeichi (s.hobeichi@unsw.edu.au)

Abstract (305 words need to be reduced to 300)

Estimating future climate change and its uncertainties relies on the analysis of a range of global climate models (GCMs) and the assessment of their spread. To meet the spatial scales required to study the local impacts of climate change, GCMs are downscaled dynamically using regional climate models or statistically using statistical methods and machine learning (ML) techniques. Due to the high computational cost involved in dynamical downscaling (DD), only a few GCMs are considered in this approach, resulting in a limited range of predictions that might not be sufficient to effectively quantify climate-related risks. Statistical downscaling (SD), on the other hand, performs downscaling at a significantly lower cost and time, but can perform poorly when extrapolated to future climates.

We introduce a hybrid downscaling framework that leverages the merits of DD and SD while overcoming the limitations of a single approach. In the new framework, a ML model is developed for each coarse gridcell to predict the subgrid distribution of the predictand of interest by combining local climate factors and subgrid land surface characteristics. The target fine-scale data needed for training ML are sourced from dynamically downscaling ten representative years.

As a proof of concept, we apply the new framework to downscale daily Evapotranspiration from BARRA-R reanalysis by a scale factor of 9 from 12.5km down to 1.5km. We use a Multilayer Perceptron and Multiple Linear Regression bias corrected with Random Forest, and we evaluate their performance at the gridcell level. Both ML models achieve high temporal and spatial efficiency out-of-sample. Additionally, through transfer-learning, it is possible to use trained models to downscale new regions at a negligible cost and with a minor performance loss.

The proposed framework offers a low-cost and fast way of downscaling a large ensemble of GCMs. Ultimately, this will enable more actionable climate projections for climate policies and adaptation planning.

1. Introduction (1248 words)

A spatially detailed understanding of climate change is critical for assessing its local impacts on water resources, agriculture, and extreme climate events (Marengo & AmBriZZi, 2006) as well as for guiding climate risk management and adaptation planning (Wilby et al., 2004). Local climate change information is typically obtained by downscaling global climate models (GCMs), which are the primary tool for simulating past and future climates under different emission scenarios (Van Vuuren et al., 2011) at large spatial scales (25 to 250 km). Due to substantial biases in GCMs (Collins et al., 2013; Raju & Kumar, 2020), it is necessary to analyse the spread across a large ensemble thereof in order to draw climate inferences (Semenov & Stratonovitch, 2010). In a similar vein, a large ensemble of downscaled GCMs needs to be employed to assess the fidelity of the projected climate change and its impacts at the local scale (Graham et al., 2007).

A variety of approaches have been developed for downscaling climate data, broadly categorized as dynamical downscaling (Giorgi, 2019; Xu et al., 2019) and statistical downscaling (Maraun & Widmann, 2018). Dynamical downscaling (DD) uses high-resolution physics-based regional climate models (RCMs; (Rummukainen, 2010)) to provide spatially and temporally detailed climate information at scales finer than those of which GCMs are capable (Giorgi, 2019; Xu et al., 2019). RCMs are forced by boundary lateral conditions from GCMs and can simulate and project the atmospheric and land surface processes that determine the regional or local terrain. The international Coordinated Regional Climate Downscaling Experiment (CORDEX; (Giorgi et al., 2009; Giorgi & Gutowski Jr, 2015)) coordinates the application and evaluation of DD through international collaboration and provides dynamically downscaled climate data derived from a suite of GCMs down to a spatial resolution of 12Km – 50km. However, due to the considerable computational cost and running times involved in DD, only a few number of GCMs are selected to force RCMs at any time period, region and/or emission scenario (Giorgi et al., 2009).As a result, for any application the available number of RCM simulations is too small to enable a meaningful characterisation of uncertainties in the downscaled variable (Hawkins & Sutton, 2009).

As for statistical downscaling, a mathematical relationship is established between the coarse output of a GCM and a fine-scale predictand of interest sourced from observations or reanalysis over the same time period and spatial domain (Benestad et al., 2008). Typically, a relationship is developed on a representative historical period and later applied on future climate projections (Vaittinada Ayar et al., 2016; Walton et al., 2020). Traditional statistical downscaling techniques include perfect prognosis (Schubert, 1998), model output statistics (MOS) with bias correction (Eden & Widmann, 2014), and stochastic weather generators (Wilks & Wilby, 1999), these are briefly described in (Gutiérrez et al., 2019). Machine learning techniques have been of increasing interest in the climate science community (Huntingford et al., 2019) and have been explored for their capabilities in statistical downscaling (Rampal et al., 2022). As a result, machine learning

for downscaling has made significant strides, especially due to the availability of a wide range of free software packages (e.g. Tensorflow (Prakash et al., 2021), Pytorch (Imambi et al., 2021), Keras (Gulli & Pal, 2017), Fastai (Howard & Gugger, 2020), and RAPIDS (RAPIDS Development Team, 2018)) as well as their ability to utilise the latest GPU technology (Z. Wang et al., 2019) which enabled fast and low-cost downscaling.

Various ML algorithms have been used for downscaling, and while they can be classified in many ways, in this context, it is reasonable to distinguish between two categories based on the shape of input predictors they can process: map-fed models and gridcell-fed models. A training sample for map-fed models consists of a stack of feature maps enabling the models to automatically extract features and incorporate spatio-temporal relationships into the fine-scale predictions. Ideally, the input domain should be large (e.g. a continent) or grid-dense (e.g. a smaller region with high original resolution). Map-fed models have been implemented in many studies including for example, Convolutional Neural Networks implementations in (Baño-Medina et al., 2021), Autoencoder architectures in (Babaousmail et al., 2021), and Generative Adversarial Networks in (Leinonen et al., 2020), typically for continental-scale downscaling. A training sample in gridcell-fed models consists of a list of feature values encoded by the user, which is usually extracted from a single gridcell or a collection of spatially coherent gridcells (Bedia et al., 2020) such as neighbouring gridcells or gridcells with common characteristics. Gridcell-fed algorithms have been applied over both small and large areas, such as Support Vector Machines in (Tripathi et al., 2006), Linear Regressions in (Sharifi et al., 2019), Random Forest (Chen et al., 2019), Generalized Linear Model in (Baño-Medina et al., 2020), Genetic Programming (Hadi Pour et al., 2014), Multilayer Perceptron (Ahmed et al., 2015), and Recurrent Neural Networks (Misra et al., 2018).

Compared to DD, the implementation of SD is very fast and less computationally intensive (J. Wang et al., 2021). However, SD models can perform poorly under extrapolation to future climates as most of them do not account for the anticipated non-stationary relationships between predictors and predictands under climate change (Hernanz et al., 2022; Hewitson et al., 2014; Lanzante et al., 2018; Salvi et al., 2016; Schoof, 2013), however, with some exceptions (Baño-Medina et al., 2022; Pichuka & Maity, 2018). Despite the fairly low cost of SD, it is only possible to implement it in regions where fine-scale data is available for training, and little is known about its ability to perform downscaling in regions outside of the training domain (F. Wang et al., 2021).

In this work, we respect the merits of each downscaling approach, and we present a hybrid downscaling framework that leverages their advantages and overcomes their limitations. By integrating DD and machine learning, we ensure that DD is implemented minimally while providing sufficient fine scale data to train ML as well as good enough information about the nonstationarity of climate. The trained ML models can then complement DD and perform downscaling at a significantly reduced cost and time as compared to implementing DD alone. We also evaluate the suitability of the hybrid downscaling approach for transfer learning, i.e., for downscaling new regions without additional ML training.

As a proof of concept, we apply this framework to downscale daily evapotranspiration (ET) from the Bureau of Meteorology Atmospheric high-resolution Regional Reanalysis for Australia (BARRA-R; (Su et al., 2019)) by a scale factor of 9. We focus on ET because it has received little attention in the literature as opposed to precipitation and temperature which have dominated most downscaling efforts. It is well known that ET provides essential information for local adaptation planning to climate change impacts on agriculture, water supply, and fire risk, among others (nice ref.?). Dynamically downscaled BARRA-R ET is obtained from BARRA-SY dataset (Su et al., 2021) and is used to build Multilayer Perceptron models, and Multiple Linear Regression models with Random Forest bias correction. We incorporate grid-fed models because they can be adapted to meet the scale requirements of a wide range of applications and end-users (e.g. applications in a farm, agricultural land or catchment) without requiring large-scale data for training, as well as being suitable for downscaling large areas similarly to their map-fed counterparts (Baño-Medina et al., 2020).

The paper is organised as follows. In the next section, we describe the training dataset, the employed machine learning models and the hybrid framework. Then, we demonstrate the temporal and spatial efficiency of the framework and its spatial extrapolation abilities through transfer learning. Finally, we conclude by discussing the significance of this work.

2. Data and Method (2034, could be reduced to 1409 if ML description is moved to supplementary)

We develop a hybrid downscaling framework that incorporates DD and SD, and we apply it to downscale daily ET by a scale factor of 9 from 12km down to 1.5km. As a result, downscaling a single gridcell produces 64 to 81 subgrid ET fields. Coarse scale ET is obtained from BARRA-R ET dataset, available for 29 years from 1990 to 2018. Fine-scale target ET is obtained from BARRA-SY which is the dynamically downscaled counterpart of BARRA-ET by the UK Met Office Unified Model over Sydney. We randomly select 20 coarse gridcells (Figure S1) on which we apply and validate the framework. The terms 'coarse' and 'fine' scales are used hereafter to refer to spatial resolutions of 12 km and 1.5 km, respectively.

The framework can be summarised in the following steps:

1. Identify 10 training years that best represent the distribution of the variable to be downscaled along the spatial (20 coarse grids) and temporal (entire timeseries) domains.
2. Perform DD in the training years to generate fine-scale data for ML training, and later use the out-of-sample years for validation.

Next, for every gridcell,

3. Train a ML model using output from DD as a target predictand, and relevant predictors that affect the spatio-temporal variation of the predictand. Here we use Multilayer Perceptron and Multiple Linear Regression with Random Forest bias correction.
4. Use the trained ML models to downscale the entire time series.

2.1 Selecting a representative sample

We identify a representative group of 10 years that best captures the long-term range and distribution of coarse ET across all gridcells. Though the length of BARRA datasets is considerably shorter than the length GCMs to which this framework is intended to be applied, the steps involved in selecting the most representative 10 years remain the same.

The training years are selected by using the *FairSubset* algorithm, described in (Ortell et al., 2019) and available as an R package. The algorithm chooses a sample from a distribution that best matches the entirety of the data by retaining the distribution information such as the median and standard deviation. Firstly, for each coarse gridcell, we compute the distributions of yearly minimum and yearly maximum ET. Next, we apply the *FairSubset* algorithm to determine the most representative five years of each distribution. We repeat this process for each distribution and coarse grid. Finally, we select five years that are most frequently observed across all representative samples in each of the minimum and maximum ET distributions. As a result, we identify 10 years that collectively best capture the lower and upper tails of ET distributions for all grid cells, and, consequently, the range of ET values that lie between them. In order to validate this method, one grid cell is excluded, and the most representative ten years are identified from the remaining 19 grid cells. Next, we ensure that this split leads to the largest overlap in probability distribution function of ET between training and out-of-sample years at the excluded gridcell (results not shown). Supplementary Figure S2 illustrates the probability density function of ET in the training sample (10 years) and the remaining years as well as the percentage overlap between them.

In this framework, dynamical downscaling needs to be performed only for the selected training years, which will provide the target predictand for ML training.

2.2 Predictor variables

As a component of the surface water, energy and carbon cycles on land, ET is influenced by meteorological factors and land surface characteristics. We use 22 predictors, consisting of 16 coarse daily meteorological variables and six fine time-invariant (static) land features including topography, and soil and vegetation properties (Figure S1). Coarse predictors include the local ET to downscale, ET from the 8 closest gridcells and a range of local meteorological variables that affect ET, listed in Table 1. The target predictand is fine-scale daily ET.

Table1: List of predictors used to downscale ET from 12 km to 1.5 km using machine learning. Coarse scale predictors are extracted from BARRA-R, and fine scale predictors are derived from BARRA-SY. Daily predictors are derived from hourly data by simple average. We note that all predictors have been standardized before training ML

Predictors	**Spatio-temporal resolution**	**Source**
Evapotranspiration (W m^{-2}) Evapotranspiration from 8 neighboring gridcells (W m^{-2})	12 km, daily	BARRA-R
Net surface shortwave radiation (W m^{-2})	12 km, daily	BARRA-R
Net surface longwave radiation	12 km, daily	BARRA-R
Air pressure at mean sea level	12 km, daily	BARRA-R
Temperature at 1.5 m (°K)	12 km, daily	BARRA-R

Specific humidity at surface	12 km, daily	BARRA-R
Wind U-component at 10m (m s^{-1})	12 km, daily	BARRA-R
Wind V-component at 10m (m s^{-1})	12 km, daily	BARRA-R
Surface albedo	1.5 km, static	BARRA-SY
Soil porosity	1.5 km, static	BARRA-SY
Soil bulk density (kg m^{-3})	1.5 km, static	BARRA-SY
Topography (m)	1.5 km, static	BARRA-SY
Leaf Area Index of plant functional type	1.5 km, static	BARRA-SY
Canopy height (m)	1.5 km, static	BARRA-SY

Collectively the selected coarse and static predictors drive the spatio-temporal variation of ET. A gridcell's ET patterns are primarily driven by its subgrid land properties, whereas its temporal variation is driven by the time-variant climate factors. ET from neighbouring gridcells were found to significantly improve downscaling, and were therefore included as predictors. This was determined by a variable importance analysis based on Random Forest feature importance (Altmann et al., 2010) that helped identify the relative contribution of each predictor to reducing downscaling error (Supplementary Figure S3).

2.3 Machine learning models

Three machine learning algorithms, MLP, MLR and MLR-RF, are trained on each coarse gridcell separately and are used to downscale ET locally (i.e. in the same gridcell they were trained on). Each algorithm is configured identically, and its tuning parameters are the same across all gridcells.

Multilayer perceptron (MLP)

Multilayer perceptron (Gardner & Dorling, 1998) is a neural network algorithm that can learn nonlinear relationships between predictors and target predictand(s). We build MLP from two hidden layers and one output layer. The first hidden layer contains 64 perceptrons, and takes as input a matrix of 22 predictors and N = 50 rows, where N is the batch size, i.e. the number of samples that enter the network in each iteration. The second layer contains 32 perceptrons and is fed by the output of the previous hidden layer. Two transformations occur at each perceptron, a linear combination of the input, followed by a rectified linear activation unit ReLu(x) = max(x,0). The activation function makes a perceptron non-linear and determines whether its output should be passed on to the following layer, i.e. if the perceptron will be activated. The output layer of MLP contains a single perceptron without an activation function and generates a value that will be compared against the corresponding target predictand by computing the loss function of the network, here, set to be the root mean square error (RMSE). After each forward pass through the network, the weights that form the linear combination at every perceptron, and other parameters of MLP are adjusted through backpropagation. During backpropagation, the expression of the loss function in terms of the current weights and other parameters, feedbacks back into the previous layers and its partial derivatives are calculated at each perceptron with respect to the perceptron's parameters (e.g. weights). The weights are updated at each iteration until the total number of epochs, here 50 epochs, is completed. The number of epochs determines how many times the network sees the entire training set. The learning rate was set to 10^{-4} and an

adaptive moment estimation algorithm known as ADAM; (Kingma & Ba, 2014) was used to adapt the learning rate for each weight of the network at each iteration during back propagation.

Multiple Linear Regression (MLR)

MLR (Maulud & Abdulazeez, 2020) establishes a linear relationship between multiple predictors and a target variable and makes prediction according to the equation $y_{MLR} = w_0 + \sum_{i=1}^{N} w_i x_i$ where y_{MLR} is the output of MLR, N= 22 is the number of predictors, $x_1 \; through \; x_{22}$ are the predictors, and $w_0 \; through \; w_{22}$ are the parameters that the model needs to tune. Finding the best parameters of the model is achieved through minimizing the discrepancy between y_{pred} and the corresponding target variable using multivariate Mean Squared Error (MSE) cost Function. Minimising the cost function is achieved through an optimization algorithm called Gradient Descent (Ruder, 2016) which involves calculating partial derivative of the expression of the loss function with respect to each parameter, then using this value and a learning rate to update the parameters of the model simultaneously. At convergence of the model to the optimal parameters, the partial derivatives are close to zero.

MLR followed by bias correction using Random Forest (MLR-RF)

To address the inability of MLR to capture nonlinear relationships between the predictors and the target variable, we incorporate RF to bias correct the fine scale ET fields derived from MLR. This involves calculating the bias of the predicted fine scale ET by subtracting the corresponding target ET $bias = y_{MLR} - y_{target}$ for the entire training sample and using it as the target variable in building RF. RF is trained to predict the bias of the fine scale ET as a function of coarse scale climate conditions and fine scale characteristics of the land bounded within the coarse grid box. These are the same predictors used as input in MLR (and MLP) and listed in Table 1. The bias terms generated by Random Forest y_{RF} are then subtracted from y_{MLR} to derive the bias corrected fine scale ET y_{MLR-RF} where $y_{MLR-RF} = y_{MLR} - y_{RF}$

2.4 Transfer-learning for spatial extrapolation

We examine the spatial transferability of the hybrid downscaling approach by evaluating the performance of trained MLP and MLR-RF models at downscaling ET in gridcells they have not been trained on. For this purpose, we downscale each coarse grid but this time using a ML model that was trained on a donor gridcell and we compare the downscaled timeseries with its dynamically downscaled counterpart.

The choice of a donor gridcell for a given recipient is made by selecting one with a similar ET distribution as the recipient's. To achieve this, we classify land into ET classes based on five predictors that summarise the distribution of ET at each coarse grid, these are the yearly minimum, maximum, first quartile, median and third quartile ET averaged along the temporal domain. We use k-mean unsupervised classifier (MacQueen, 1967) based on an algorithm developed by (Lloyd, 1982) resulting in 11 classes. The optimal number of classes, i.e. 11, is chosen using a graphical silhouettes analysis (Rousseeuw, 1986). As a result, each coarse gridcell

is assigned a label consisting of an alphabet that refers to its ET class and an index that distinguishes it from other gridcells in the same class (Figure S1). Subsequently, gridcells of the same class, such as A_1, A_2 and A_3, may be considered as appropriate donors for each other. Among all potential donors of a recipient gridcell, we pick one with the closest land features using dissimilarity in static predictors as proxy for distance, such that it minimizes:

$$dissimilarity_{donor,recepient} = \sqrt{\sum_{predictor}^{LAI,\ soil\ albedo\ ,\ topography} \sum_{q=1}^{3} \left(quartile_q\left(predictor_{donor}\right) - \ quartile_q\left(predi\right.\right.}$$

Where *predictor* ∈ {soil albedo, LAI, and topography} and *q* refers to a quartile value, i.e. $quartile_1$, $quartile_2$, and $quartile_3$ are the first quartile, median and third quartile respectively.

For example, to select the best donor for A_3 among A_1 and A_2, we calculate $dissimilarity_{A_3,A_1}$ and $dissimilarity_{A_2,A_1}$, then we select the donor that achieve the minimum thereof.

We confirm that selecting a donor gridcell this way is effective by verifying that none of the ML model trained on the remaining 18 coarse grids is able to outperform the one trained on the selected donor (results not shown).

2.5 Metrics of performance

To evaluate the framework, we examine the temporal and the spatial efficiencies of MLP and MLR-RF in each of the 20 gridcells and out-of-sample, i.e. in years that were not included in the training. Performance is determined by how well these ML models are able to emulate DD.

Temporal efficiency is measured by temporal correlation, percentage of bias relative to mean computed along the temporal domain, the coefficient of determination (R^2; (Di Bucchianico, 2008)), and Kling-Gupta Efficiency (KGE; (Pool et al., 2018)). The best score is zero for percentage of bias and *one* for the remaining metrics. What the literature considers as good KGE score varies from *greater or equal to -0.41*(Knoben et al., 2019) to *greater or equal to 0.75* (Towner et al., 2019). These metrics are calculated yearly for each fine grid, and summarized in a boxplot per coarse grid for each metric.

Spatial efficiency is measured by histogram match (Swain & Ballard, 1991) which calculates the degree to which the spatial pattern of ET within a coarse grid predicted by ML matches its DD counterpart. Histogram match (γ) is calculated as the histogram intersection for ML and DD, each containing *n* bins, where *n* is the square root of the number of simulated fine gridcells. γ score is the proportion of gridcells from the ML histogram that have corresponding gridcell of the same bin in the DD histogram. Daily γ is calculated out-of-sample and the results are summarized in a boxplot per coarse grid.

Lastly, we illustrate the spatial pattern of the climatological mean and bias ET predicted by ML for every gridcell to enable a visual assessment of the ML predictions at the subgrid scale.

3. Results (682 words)

3. 1 Temporal efficiency

Figure 1 illustrates the out-of-sample performance of the ML models for all 20 coarse gridcells. As explained earlier, for each coarse gridcell, boxplots summarise the scores achieved by MLP or MLR-RF in each of the 64-81 subgrid calculated for the out-of-sample years. MLR is included in the plots to highlight the improvement that bias correction with RF brings to MLR. Overall, the KGE scores for MLP and MLR-RF lie between 0.8 and 0.9 in the majority of gridcells, and their correlation and R^2 scores are greater than 0.8 and 0.7 respectively. The performance of both ML models is slightly reduced in coarse gridcell $C_{3,}$ which can be explained by a lower overlap (93%) of training data with the out-of-sample (testing) data in comparison to *larger than 95% overlap* in all the other gridcells as shown in Supplementary Figure S2. MLP and MLR-RF biases lie between -5% and 5% across all gridcells.

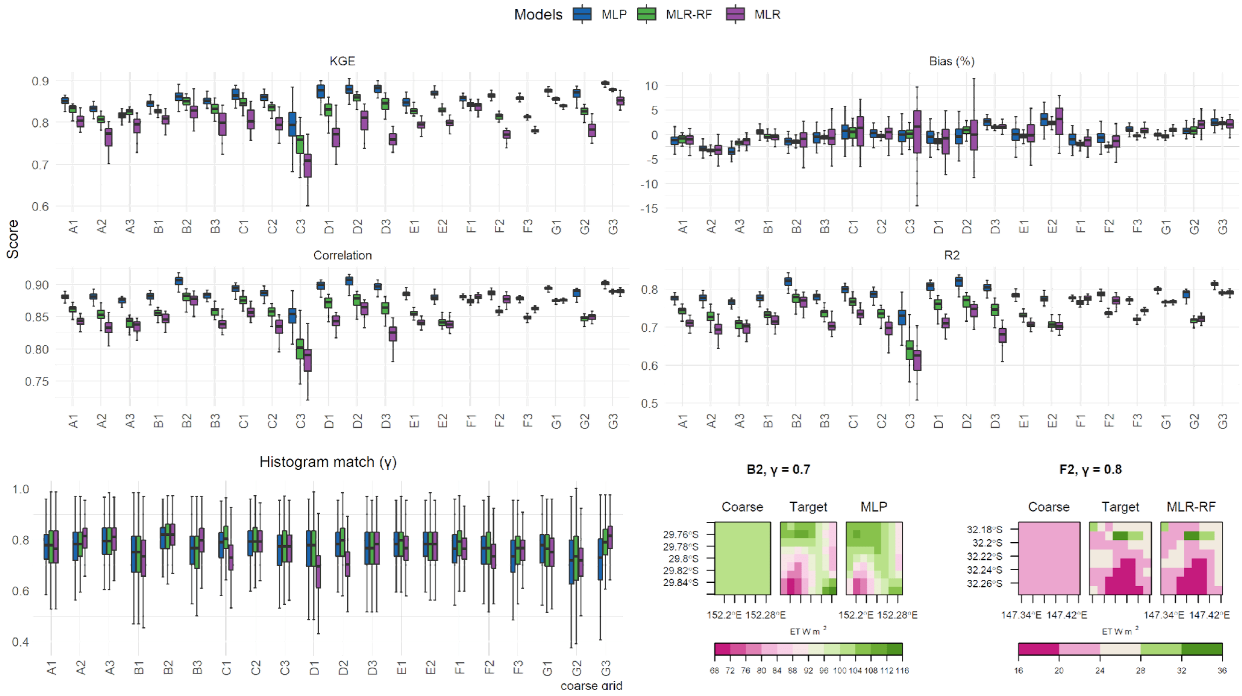

Figure 1: Out-of-sample performance of MLP, MLR-RF and MLR at 20 coarse gridcells across five metrics of performance, KGE, bias relative to temporal mean (Bias %), correlation, coefficient of determination R^2, and histogram match (see *Section 2.5* for a description of each metric). The grid boxes illustrate 2 examples where histogram match scores 0.7 and 0.8 for MLP and MLR-RF respectively.

3.2 Spatial efficiency

The histogram match score (γ) is calculated daily for the out-of-sample years and summarized in a boxplot for each coarse gridcell. MLP and MLR-RF score between 0.7 and 0.8 for the majority of predictions. For a better understanding of how the spatial match in reality looks at these values, we provide a visual representation of subgrid ET patterns derived by ML and its target counterpart in two scenarios in which the scores achieved are 0.7 and 0.8.

Figure 2 illustrates the spatial patterns of climatological mean ET computed for MLR-RF, MLP, and DD as well as climatological bias (i.e., ML – DD). A visual assessment of the results indicates that the subgrid patterns of ET predicted by the ML models are highly consistent with those derived from DD. Additionally, at any coarse gridcell, instances where the magnitude of climatological bias exceeds 2 Wm^{-2} occur at most in two fine grids out of the 64 - 81 subgrids.

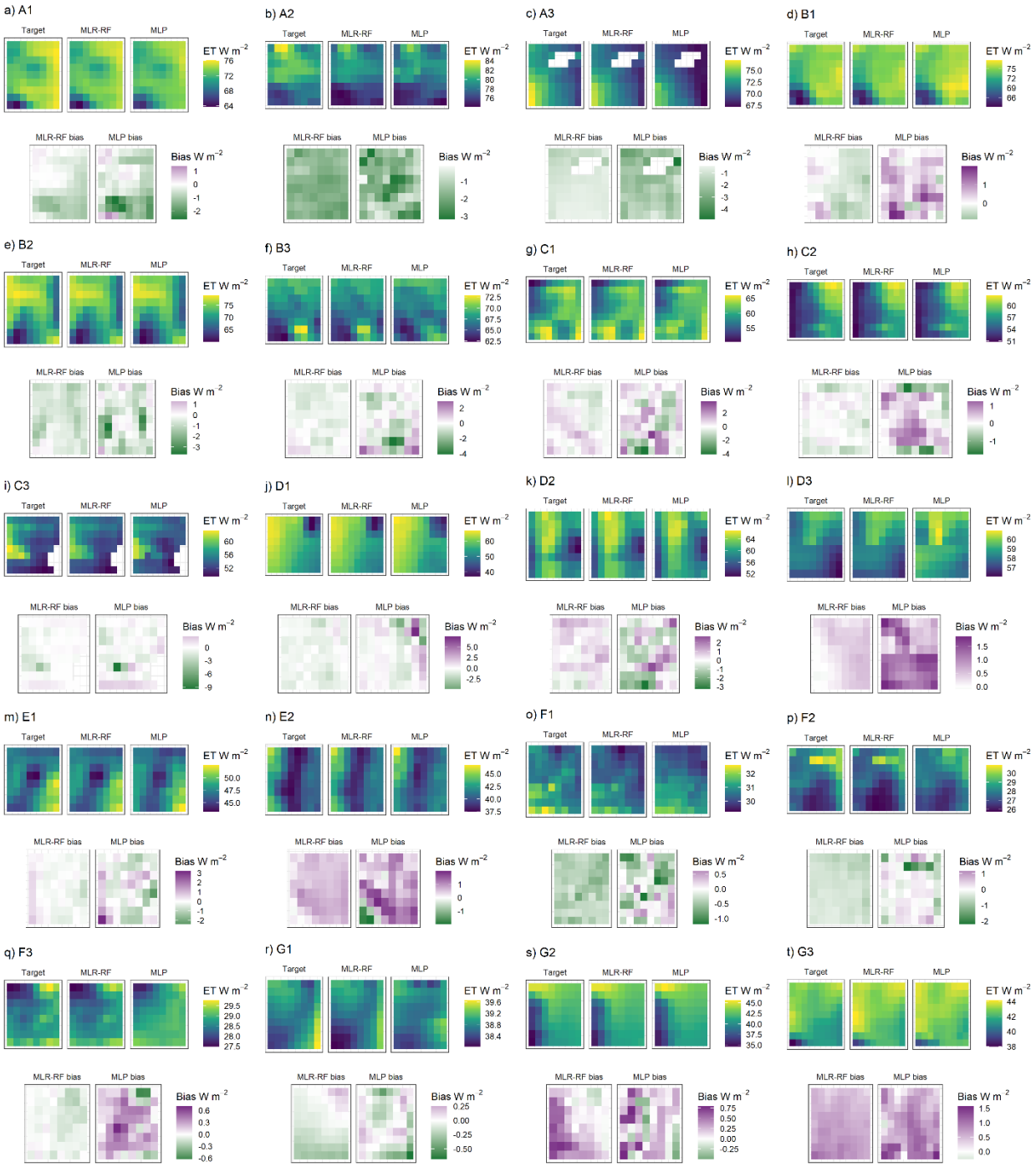

Figure 2: Subgrid patterns of climatological mean ET computed for MLR-RF, MLP, and DD (Target) and climatological bias (i.e. ML – DD) exerted by MLP and MLR-RF.

3.3 Transfer learning

For each coarse gridcell, we identify the best donor among the remaining 19 gridcells first by identifying all gridcells who are within the same ET class (potential donors), then choosing among those the one whose land surface characteristics are the closest to the recipient gridcell, i.e. the one that minimizes the dissimilarity equation presented in *Transfer learning for spatial extrapolation*. ML models transferred from a donor gridcell are denoted tMLP and tMLR-RF, and their performance is presented in Figure 3. At 80% of gridcells, the median KGE scores for tMLP and tMLR-RF lie between 0.7 and 0.9, their correlation and median R^2 scores are greater than 0.8 and 0.7 respectively, and their median relative biases lie between -10% and 10%. Similarly, the median spatial efficiency is greater than 0.7 across 80% of gridcells.

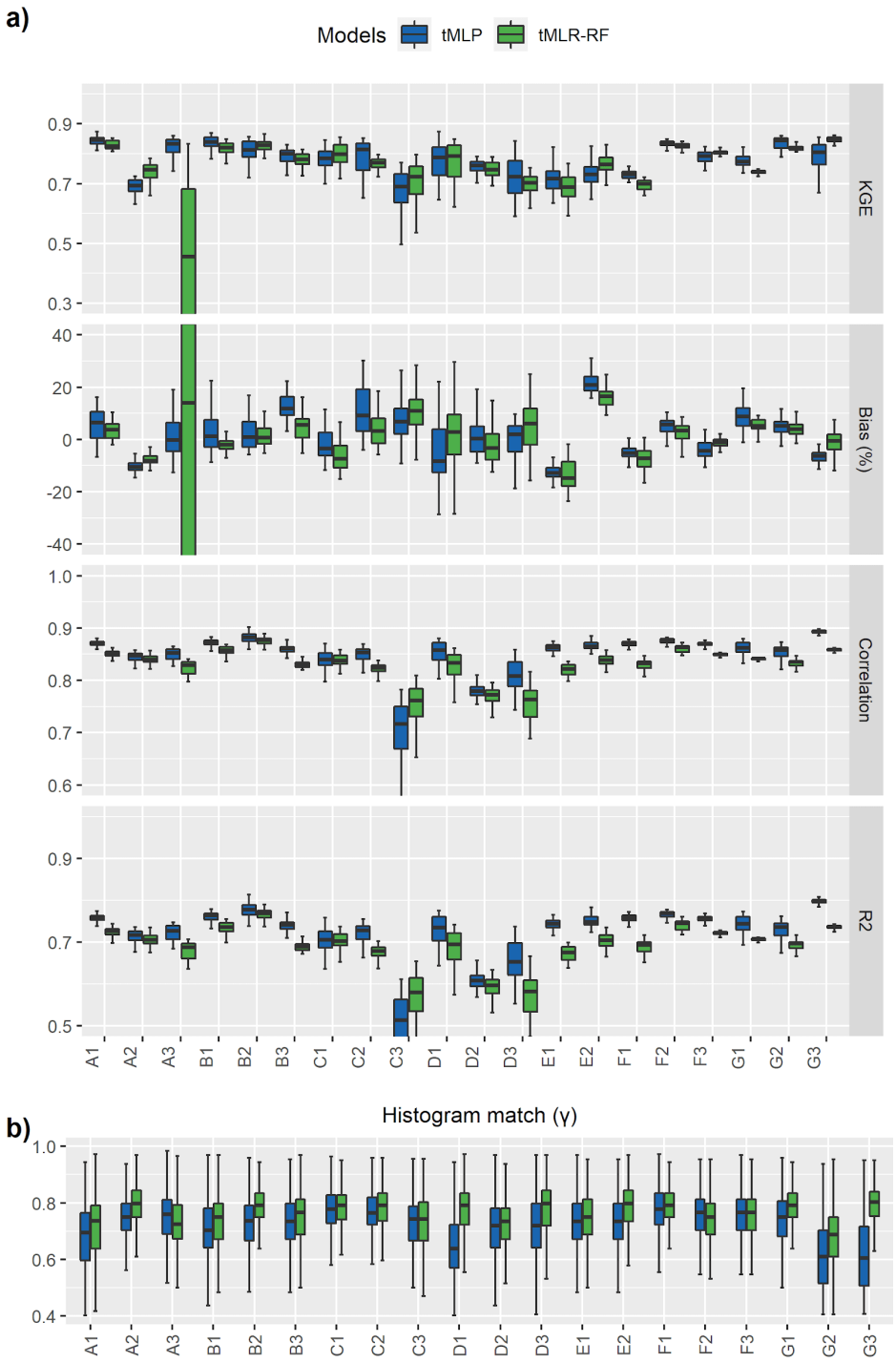

Figure 3: Out-of-sample performance of tMLP, and tMLR-RF at 20 coarse gridcells across (a) temporal efficiency metrics of performance KGE, bias relative to temporal mean (Bias %), correlation, coefficient of determination R^2, and (b) spatial efficiency metric of performance, histogram match (see *Section2.5* for a description of each metric). tMLP (tMLR-RF) refers to MLP (MLR-RF) trained on a donor gridcell.

Figure 4 illustrates the spatial patterns of subgrid climatological bias as a percentage of the mean target. It also shows the probability distribution of daily ET at each coarse gridcell (recipient) and its donor. We compute for every coarse grid the median absolute bias (%) as well as its upper quartile along the 64-81 subgrids. For instance, a median absolute bias of 10% indicates that in half of the fine gridcells derived by downscaling a given coarse gridcell, ML bias is less than 10%. We find that in 80% of the grid cells both tMLP and tMLR-RF achieve a median bias less than 10% and an upper quartile bias less than 15%. Larger biases are exhibited in A_3 by tMLR-RF when trained on A_1 which might be due to the subtantial dissimilarity in subgrid topography between the donor gridcell (ranging from 150 to 400m) compared to less than 40m across A1, which does not occur between any other donor-recipient gridcells. Although this difference is significant, it does not induce large biases in tMLP.

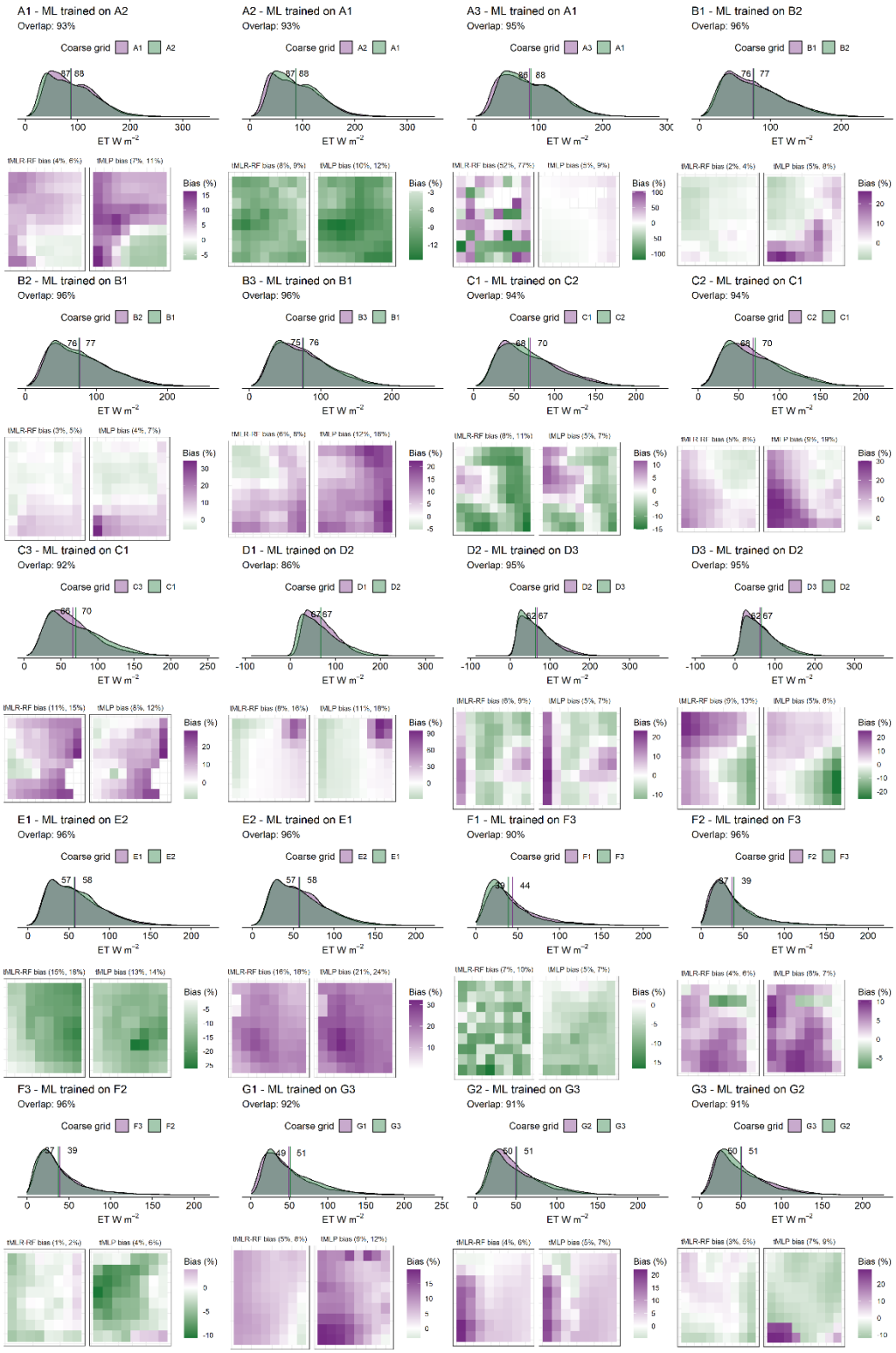

Figure 4: Relative probability densities of coarse ET for a coarse gridcell (purple) and its donor (green) and their overlap (grey). The vertical purple and green lines indicate the median values for the recipient and donor gridcells respectively. The gridboxes illustrate the subgrid patterns of climatological bias on the recipient gridcells induced by tMLP and tMLR-RF which have been trained on the donor gridcell.

This hybrid downscaling approach has proven to be efficient in retaining both the spatial and temporal characteristics of the target predictand. Although the performance of the trained MLP and MLR-RF is reduced when transferred to new grid cells, the fact that the downscaling costs are negligible, since no dynamical downscaling or ML training is required, makes this approach quite valuable.

4. Discussion and Conclusion (602)

We present a hybrid downscaling framework that combines statistical and dynamical downscaling and overcomes the limitations of a single approach. Compared to applying dynamical downscaling alone, hybrid downscaling cuts computational costs and runtimes significantly by restricting the implementation of DD to 10 years only, followed by a speedy downscaling using ML over the entire time range. By ensuring that the training years are representative, plausible future change would fall within the calibration range of ML, thereby addressing the well-known issue of poor performance when machine learning downscaling is applied to future projections (Hernanz et al., 2022; Lanzante et al., 2018).

With this framework, multiple GCMs could be downscaled at the same cost as dynamically downscaling a single one, thereby enabling cost- and time-efficient assessment of a range of plausible climate projections. Also, the hybrid framework can help fill temporal gaps in the available RCM simulations from the CORDEX framework, thus enabling a more comprehensive assessment of climate change at the local scale. Further, this framework can be applied at a specific location, even at the scale of a single gridcell, such as a farm, agricultural land, or catchment, benefiting a variety of users and applications over a range of spatial extents.

A further advantage of the proposed approach is that it enables downscaling regions outside the spatial domain of RCMs through transfer-learning, i.e., by using ML downscaling models that have been trained elsewhere. While downscaling performance may be reduced under transfer-learning, the cost of downscaling is negligible in this case, which makes transfer-learning beneficial.

This framework can be adapted to other climate variables such as land surface fluxes (sensible heat flux, ground heat flux, and runoff) and atmospheric variables (e.g., precipitation P, temperature T, and wind). Relevant coarse and static predictors may be the same as those used here when the predictand of interest is sensible heat flux and ground heat flux. Additional static predictors providing hydrography information (e.g. orography and slopes) are needed for downscaling runoff. Atmospheric variables are heavily influenced by non-local processes that need to be captured by the predictors space for successful downscaling. This can be addressed by including information from neighbouring grid cells or spatially coherent regions where key atmospheric processes occur. Map-fed ML models employ convolutional layers that automatically extract the important features from relevant geographic domain (Rampal et al., 2022), thereby reducing the need for manual selection of predictors. Map-fed models have therefore become increasingly popular for downscaling P and T over large geographical areas and have shown promising results (Baño-Medina et al., 2021, 2022; Rampal et al., 2022). Yet,

using map-fed models may be less beneficial, or even problematic for downscaling ET and other land surface fluxes. The reason is that these models need to be fed with large rectangle-shaped images (maps) which typically cover parts of the ocean, however, in many cases, static variables relevant to land surface fluxes are irrelevant over non-land, thus introducing missing values that convolutional layers cannot handle. Including time-lag climate variables as predictors can further improve downscaling by capturing spatio-temporal dependencies (Misra et al., 2018). This can be done manually by including climate variables from previous time step(s) as additional predictor(s), or by using Recurrent Neural Networks which are designed to automatically incorporate information from previous steps. LSTM in particular has shown great potential in climate downscaling (Chou et al., 2021; Li et al., 2020; Mouatadid et al., 2017; Tran Anh et al., 2019). This ML algorithm could be explored in the future within this framework as an alternative to MLP and MLR-RF.

A limitation of this framework is that at best it can perform similarly to DD.

5. Author contribution

S.H. and N.N. designed the study with advice from G.A., A.P., S.S., C.B., and Y.S.. S.H. performed the analysis and wrote the script with technical assistance from S.G.. S.H. interpreted and discussed the results with input from A.P., G.A., N.N., Y.S., C.B. and S.S. All authors reviewed and edited the manuscript before submission.

6. Data availability

This study uses BARRA-R and BARRA-SY datasets which are freely available for academic use on the Australian National Computational Infrastructure (NCI). Data access is available upon request to the data developers.

7. Code availability and computational efficiency

The ML downscaling scripts will be available as part of a GitHub repository once the paper has been accepted.

It took ~xmin to train a single grid and a few seconds to predict the subgrid daily ET over the entire temporal domain 1990-2018. Using parallel computing with xx CPUs, it is possible to downscale xxx gridcells in xxx minutes. We use free software packages including FastAI, PyTorch, NVIDIA GPU API with x GPU– to be updated.

8. Acknowledgment

All the authors acknowledge the support of the Australian Research Council Centre of Excellence for Climate Extremes (CLEX; CE170100023). This research was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government. S.H. thanks members of the Computational Modelling Systems team at CLEX for providing computational support on NCI supercomputers.

9. References

Ahmed, K., Shahid, S., Haroon, S. Bin, & Xiao-Jun, W. (2015). Multilayer perceptron neural

network for downscaling rainfall in arid region: a case study of Baluchistan, Pakistan. *Journal of Earth System Science*, *124*(6), 1325–1341. https://doi.org/https://doi.org/10.1007/s12040-015-0602-9

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. https://doi.org/https://doi.org/10.1093/bioinformatics/btq134

Babaousmail, H., Hou, R., Gnitou, G. T., & Ayugi, B. (2021). Novel statistical downscaling emulator for precipitation projections using deep Convolutional Autoencoder over Northern Africa. *Journal of Atmospheric and Solar-Terrestrial Physics*, *218*, 105614. https://doi.org/https://doi.org/10.1016/j.jastp.2021.105614

Baño-Medina, J., Manzanas, R., & Gutiérrez, J. M. (2020). Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, *13*(4), 2109–2124. https://doi.org/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*(11–12), 2941–2951. https://doi.org/10.1007/s00382-021-05847-0

Baño-Medina, J., Manzanas, R., Cimadevilla, E., Fernández, J., González-Abad, J., Cofiño, A. S., & Gutiérrez, J. M. (2022). Downscaling multi-model climate projection ensembles with deep learning (DeepESD): contribution to CORDEX EUR-44. *Geoscientific Model Development*, *15*(17), 6747–6758.

Bedia, J., Baño-Medina, J., Legasa, M. N., Iturbide, M., Manzanas, R., Herrera, S., et al. (2020). Statistical downscaling with the downscaleR package (v3. 1.0): contribution to the VALUE intercomparison experiment. *Geoscientific Model Development*, *13*(3), 1711–1735. https://doi.org/https://doi.org/10.5194/gmd-13-1711-2020

Benestad, R. E., Chen, D., & Hanssen-Bauer, I. (2008). *Empirical-statistical downscaling*. World Scientific Publishing Company. https://doi.org/https://doi.org/10.1142/6908

Di Bucchianico, A. (2008). Coefficient of determination (R 2). *Encyclopedia of Statistics in Quality and Reliability*, *1*. https://doi.org/https://doi.org/10.1002/9780470061572

Chen, L., He, Q., Liu, K., Li, J., & Jing, C. (2019). Downscaling of GRACE-derived groundwater storage based on the random forest model. *Remote Sensing*, *11*(24), 2979. https://doi.org/https://doi.org/10.3390/rs11242979

Chou, C., Park, J., & Chou, E. (2021). Generating High-Resolution Climate Change Projections Using Super-Resolution Convolutional LSTM Neural Networks. In *2021 13th International Conference on Advanced Computational Intelligence (ICACI)* (pp. 293–298). IEEE. https://doi.org/https://doi.org/10.1109/ICACI52617.2021.9435890

Collins, M., Knutti, R., Arblaster, J., Dufresne, J.-L., Fichefet, T., Friedlingstein, P., et al. (2013). Long-term climate change: projections, commitments and irreversibility. In *Climate change 2013-The physical science basis: Contribution of working group I to the fifth assessment report of the intergovernmental panel on climate change* (pp. 1029–1136). Cambridge

University Press.

Eden, J. M., & Widmann, M. (2014). Downscaling of GCM-simulated precipitation using model output statistics. *Journal of Climate*, *27*(1), 312–324. https://doi.org/https://doi.org/10.1175/JCLI-D-13-00063.1

Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences. *Atmospheric Environment*, *32*(14–15), 2627–2636. https://doi.org/https://doi.org/10.1016/S1352-2310(97)00447-0

Giorgi, F. (2019). Thirty Years of Regional Climate Modeling: Where Are We and Where Are We Going next? *Journal of Geophysical Research: Atmospheres*, *124*(11), 5696–5723. https://doi.org/10.1029/2018JD030094

Giorgi, F., & Gutowski Jr, W. J. (2015). Regional dynamical downscaling and the CORDEX initiative. *Annual Review of Environment and Resources*, *40*, 467–490. https://doi.org/https://doi.org/10.1146/annurev-environ-102014-021217

Giorgi, F., Jones, C., & Asrar, G. R. (2009). Addressing climate information needs at the regional level: the CORDEX framework. *World Meteorological Organization (WMO) Bulletin*, *58*(3), 175.

Graham, L., Andréasson, J., & Carlsson, B. (2007). Assessing climate change impacts on hydrology from an ensemble of regional climate models, model scales and linking methods–a case study on the Lule River basin. *Climatic Change*, *81*(1), 293–307. https://doi.org/https://doi.org/10.1007/s10584-006-9215-2

Gulli, A., & Pal, S. (2017). *Deep learning with Keras*. Packt Publishing Ltd.

Gutiérrez, J. M., Maraun, D., Widmann, M., Huth, R., Hertig, E., Benestad, R., et al. (2019). An intercomparison of a large ensemble of statistical downscaling methods over Europe: Results from the VALUE perfect predictor cross-validation experiment. *International Journal of Climatology*, *39*(9), 3750–3785. https://doi.org/10.1002/joc.5462

Hadi Pour, S., Harun, S. Bin, & Shahid, S. (2014). Genetic programming for the downscaling of extreme rainfall events on the East Coast of Peninsular Malaysia. *Atmosphere*, *5*(4), 914–936. https://doi.org/https://doi.org/10.3390/atmos5040914

Hawkins, E., & Sutton, R. (2009). The potential to narrow uncertainty in regional climate predictions. *Bulletin of the American Meteorological Society*, *90*(8), 1095–1108. https://doi.org/https://doi.org/10.1175/2009BAMS2607.1

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*, (February), 1–9. https://doi.org/10.1002/asl.1087

Hewitson, B. C., Daron, J., Crane, R. G., Zermoglio, M. F., & Jack, C. (2014). Interrogating empirical-statistical downscaling. *Climatic Change*, *122*(4), 539–554. https://doi.org/https://doi.org/10.1007/s10584-013-1021-z

Howard, J., & Gugger, S. (2020). Fastai: a layered API for deep learning. *Information*, *11*(2), 108. https://doi.org/https://doi.org/10.3390/info11020108

Huntingford, C., Jeffers, E. S., Bonsall, M. B., Christensen, H. M., Lees, T., & Yang, H. (2019). Machine learning and artificial intelligence to aid climate change research and preparedness. *Environmental Research Letters*, *14*(12), 124007. https://doi.org/10.1088/1748-9326/ab4e55

Imambi, S., Prakash, K. B., & Kanagachidambaresan, G. R. (2021). PyTorch. In *Programming with TensorFlow* (pp. 87–104). Springer. https://doi.org/https://doi.org/10.1007/978-3-030-57077-4_10

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *ArXiv Preprint ArXiv:1412.6980*. https://doi.org/https://doi.org/10.48550/arXiv.1412.6980

Knoben, W. J. M., Freer, J. E., & Woods, R. A. (2019). Inherent benchmark or not? Comparing Nash–Sutcliffe and Kling–Gupta efficiency scores. *Hydrology and Earth System Sciences*, *23*(10), 4323–4331. https://doi.org/https://doi.org/10.5194/hess-23-4323-2019

Lanzante, J. R., Dixon, K. W., Nath, M. J., Whitlock, C. E., & Adams-Smith, D. (2018). Some pitfalls in statistical downscaling of future climate. *Bulletin of the American Meteorological Society*, *99*(4), 791–803. https://doi.org/https://doi.org/10.1175/BAMS-D-17-0046.1

Leinonen, J., Nerini, D., & Berne, A. (2020). 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/https://doi.org/10.1109/TGRS.2020.3032790

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*(1), 571–588. https://doi.org/https://doi.org/10.1007/s00704-020-03098-3

Lloyd, S. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory*, *28*(2), 129–137. https://doi.org/https://doi.org/10.1109/TIT.1982.1056489

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 5.1, pp. 281–297). Oakland, CA, USA.

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

Marengo, J. A., & AmBriZZi, T. (2006). Use of regional climate models in impacts assessments and adaptations studies from continental to regional and local scales. In *Proceedings of* (Vol. 8, pp. 291–296).

Maulud, D., & Abdulazeez, A. M. (2020). A review on linear regression comprehensive in machine learning. *Journal of Applied Science and Technology Trends*, *1*(4), 140–147. https://doi.org/https://doi.org/10.38094/jastt1457

Misra, S., Sarkar, S., & Mitra, P. (2018). Statistical downscaling of precipitation using long

short-term memory recurrent neural networks. *Theoretical and Applied Climatology*, *134*(3), 1179–1196. https://doi.org/https://doi.org/10.1007/s00704-017-2307-2

Mouatadid, S., Easterbrook, S., & Erler, A. R. (2017). A machine learning approach to non-uniform spatial downscaling of climate variables. In *2017 IEEE international conference on data mining workshops (ICDMW)* (pp. 332–341). IEEE. https://doi.org/https://doi.org/10.1109/ICDMW.2017.49

Ortell, K. K., Switonski, P. M., & Delaney, J. R. (2019). FairSubset: A tool to choose representative subsets of data for use with replicates or groups of different sample sizes. *Journal of Biological Methods*, *6*(3). https://doi.org/https://doi.org/10.14440/jbm.2019.299

Pichuka, S., & Maity, R. (2018). Development of a time‐varying downscaling model considering non‐stationarity using a Bayesian approach. *International Journal of Climatology*, *38*(7), 3157–3176. https://doi.org/https://doi.org/10.1002/joc.5491

Pool, S., Vis, M., & Seibert, J. (2018). Evaluating model performance: towards a non-parametric variant of the Kling-Gupta efficiency. *Hydrological Sciences Journal*, *63*(13–14), 1941–1953. https://doi.org/https://doi.org/10.1080/02626667.2018.1552002

Prakash, K. B., Ruwali, A., & Kanagachidambaresan, G. R. (2021). Introduction to tensorflow package. In *Programming with TensorFlow* (pp. 1–4). Springer. https://doi.org/https://doi.org/10.1007/978-3-030-57077-4_1

Raju, K. S., & Kumar, D. N. (2020). Review of approaches for selection and ensembling of GCMs. *Journal of Water and Climate Change*, *11*(3), 577–599. https://doi.org/https://doi.org/10.2166/wcc.2020.128

Rampal, N., Gibson, P. B., Sood, A., Stuart, Stephen, Fauchereau, N. C., Brandolino, C., et al. (2022). High-resolution downscaling with interpretable deep learning: rainfall extremes over New Zealand. *Weather and Climate Extremes*, In Review.

RAPIDS Development Team. (2018). RAPIDS: Collection of Libraries for End to End GPU Data Science.

Rousseeuw, P. (1986). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 53–65.

Ruder, S. (2016). An overview of gradient descent optimization algorithms. *ArXiv Preprint ArXiv:1609.04747*. https://doi.org/https://doi.org/10.48550/arXiv.1609.04747

Rummukainen, M. (2010). State‐of‐the‐art with regional climate models. *Wiley Interdisciplinary Reviews: Climate Change*, *1*(1), 82–96. https://doi.org/https://doi.org/10.1002/wcc.8

Salvi, K., Ghosh, S., & Ganguly, A. R. (2016). Credibility of statistical downscaling under nonstationary climate. *Climate Dynamics*, *46*(5), 1991–2023. https://doi.org/https://doi.org/10.1007/s00382-015-2688-9

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

Schubert, S. (1998). Downscaling local extreme temperature changes in south‐eastern Australia from the CSIRO Mark2 GCM. *International Journal of Climatology: A Journal of the Royal Meteorological Society*, *18*(13), 1419–1438. https://doi.org/https://doi.org/10.1002/(SICI)1097-0088(19981115)18:13<1419::AID-JOC314>3.0.CO;2-Z

Semenov, M. A., & Stratonovitch, P. (2010). Use of multi-model ensembles from global climate models for assessment of climate change impacts. *Climate Research*, *41*(1), 1–14. https://doi.org/https://doi.org/10.3354/cr00836

Sharifi, E., Saghafian, B., & Steinacker, R. (2019). Downscaling satellite precipitation estimates with multiple linear regression, artificial neural networks, and spline interpolation techniques. *Journal of Geophysical Research: Atmospheres*, *124*(2), 789–805. https://doi.org/https://doi.org/10.1029/2018JD028795

Su, C.-H., Eizenberg, N., Steinle, P., Jakob, D., Fox-Hughes, P., White, C. J., et al. (2019). BARRA v1. 0: the Bureau of Meteorology atmospheric high-resolution regional reanalysis for Australia. *Geoscientific Model Development*, *12*(5), 2049–2068. https://doi.org/https://doi.org/10.5194/gmd-12-2049-2019

Su, C.-H., Eizenberg, N., Jakob, D., Fox-Hughes, P., Steinle, P., White, C. J., & Franklin, C. (2021). BARRA v1. 0: kilometre-scale downscaling of an Australian regional atmospheric reanalysis over four midlatitude domains. *Geoscientific Model Development*, *14*(7), 4357–4378. https://doi.org/https://doi.org/10.5194/gmd-14-4357-2021

Swain, M. J., & Ballard, D. H. (1991). Color indexing. *International Journal of Computer Vision*, *7*(1), 11–32.

Towner, J., Cloke, H. L., Zsoter, E., Flamig, Z., Hoch, J. M., Bazo, J., et al. (2019). Assessing the performance of global hydrological models for capturing peak river flows in the Amazon basin. *Hydrology and Earth System Sciences*, *23*(7), 3057–3080. https://doi.org/https://doi.org/10.5194/hess-23-3057-2019

Tran Anh, D., Van, S. P., Dang, T. D., & Hoang, L. P. (2019). Downscaling rainfall using deep learning long short‐term memory and feedforward neural network. *International Journal of Climatology*, *39*(10), 4170–4188. https://doi.org/https://doi.org/10.1002/joc.6066

Tripathi, S., Srinivas, V. V, & Nanjundiah, R. S. (2006). Downscaling of precipitation for climate change scenarios: a support vector machine approach. *Journal of Hydrology*, *330*(3–4), 621–640. https://doi.org/https://doi.org/10.1016/j.jhydrol.2006.04.030

Vaittinada Ayar, P., Vrac, M., Bastin, S., Carreau, J., Déqué, M., & Gallardo, C. (2016). Intercomparison of statistical and dynamical downscaling models under the EURO- and MED-CORDEX initiative framework: present climate evaluations. *Climate Dynamics*, *46*(3–4), 1301–1329. https://doi.org/10.1007/s00382-015-2647-5

Van Vuuren, D. P., Edmonds, J., Kainuma, M., Riahi, K., Thomson, A., Hibbard, K., et al. (2011). The representative concentration pathways: an overview. *Climatic Change*, *109*(1), 5–31. https://doi.org/https://doi.org/10.1007/s10584-011-0148-z

Walton, D., Berg, N., Pierce, D., Maurer, E., Hall, A., Lin, Y., et al. (2020). Understanding

differences in California climate projections produced by dynamical and statistical downscaling. *Journal of Geophysical Research: Atmospheres*, *125*(19), e2020JD032812. https://doi.org/https://doi.org/10.1029/2020JD032812

Wang, F., Tian, D., Lowe, L., Kalin, L., & Lehrter, J. (2021). Deep Learning for Daily Precipitation and Temperature Downscaling. *Water Resources Research*, *57*(4), 1–21. https://doi.org/10.1029/2020WR029308

Wang, J., Liu, Z., Foster, I., Chang, W., Kettimuthu, R., & Kotamarthi, V. R. (2021). Fast and accurate learned multiresolution dynamical downscaling for precipitation. *Geoscientific Model Development*, *14*(10), 6355–6372.

Wang, Z., Liu, K., Li, J., Zhu, Y., & Zhang, Y. (2019). Various frameworks and libraries of machine learning and deep learning: a survey. *Archives of Computational Methods in Engineering*, 1–24. https://doi.org/https://doi.org/10.1007/s11831-018-09312-w

Wilby, R. L., Charles, S. P., Zorita, E., Timbal, B., Whetton, P., & Mearns, L. O. (2004). Guidelines for use of climate scenarios developed from statistical downscaling methods. *Supporting Material of the Intergovernmental Panel on Climate Change, Available from the DDC of IPCC TGCIA*, *27*.

Wilks, D. S., & Wilby, R. L. (1999). The weather generation game: a review of stochastic weather models. *Progress in Physical Geography*, *23*(3), 329–357. https://doi.org/http://dx.doi.org/10.1177/030913339902300302

Xu, Z., Han, Y., & Yang, Z. (2019). Dynamical downscaling of regional climate: A review of methods and limitations. *Science China Earth Sciences*, *62*(2), 365–375. https://doi.org/10.1007/s11430-018-9261-5