Weather and Climate Extremes 38 (2022) 100525

Contents lists available at ScienceDirect

Weather and Climate Extremes

journal homepage: www.elsevier.com/locate/wace

High-resolution downscaling with interpretable deep learning: Rainfall extremes over New Zealand

Neelesh Rampal [a], Peter B. Gibson [b,*], Abha Sood [b], Stephen Stuart [b], Nicolas C. Fauchereau [a], Chris Brandolino [a], Ben Noll [a], Tristan Meyers [b]

[a] National Institute of Water and Atmospheric Research (NIWA), Auckland, New Zealand
[b] National Institute of Water and Atmospheric Research (NIWA), Wellington, New Zealand

ARTICLE INFO

Keywords:
Statistical downscaling
Deep learning
Machine learning
Precipitation extremes

ABSTRACT

The gap in resolution between existing global climate model output and that sought by decision-makers drives an ongoing need for climate downscaling. Here we test the extent to which developments in deep learning can outperform existing statistical approaches for downscaling historical rainfall in the highly complex terrain setting of New Zealand. While deep learning removes the need for manual feature selection when extracting spatial-temporal information from predictor fields, several key considerations need to be addressed. These include: the chosen complexity of the network architecture, suitable loss functions tailored to the problem, as well as input data considerations of domain size and amount of training data required to provide adequate out-of-sample generalization. Sensitivity testing to these considerations reveals that a relatively simple convolutional neural network (CNN) architecture with carefully selected loss functions can considerably outperform existing statistical downscaling models based on multiple linear regression with manual feature selection. When aggregated across the entire region, the fraction of explained variance on wet days increased from 0.35 to 0.52, the root-mean-squared error reduced by over 20% and percentage biases for the 90th percentile of rainfall improved by over 25%. Using interpretable machine learning methods, we demonstrate that the CNN has been capable of self-learning physically plausible relationships between the large-scale atmospheric environment and extreme localized rainfall events. The historical performance and physical interpretability documented here lends support for wider development and application of deep learning in climate downscaling.

1. Introduction

A major and longstanding challenge for climate science is to provide reliable and actionable climate change projections at high spatial resolution. These high-resolution projections are critical for informing localized climate change adaptation measures and preparedness for possible future changes in extreme events. Despite ongoing investment and development in global climate models (GCMs), their typical grid resolution (~100 km) still limits their direct application to local scales (Marotzke et al., 2017; Fiedler et al., 2021). Limited-area dynamical model projections, such as those coordinated through international efforts like the Coordinated Regional climate Downscaling Experiment (CORDEX; Giorgi et al., 2009), continue to play an important role in bridging this resolution divide. Nevertheless, in practice, these regional climate models (RCMs) can also be very computationally expensive, especially when run at high spatial resolution, meaning that typically only a small number of GCM/RCM pairs are used in regional projections (e.g. Jacob et al., 2014; Ruti et al., 2016; Gibson et al., 2019). Consequently, RCM-based projections risk underrepresenting both initial condition uncertainty (Xie et al., 2015; Deser, 2020) and structural model-based uncertainty across the subset of GCMs/RCMs used for downscaling.

The use of statistical downscaling has a long history of complementing dynamically based RCM projections (Wilby and Wigley, 1997; Wilby et al., 1998; Benestad, 2004; Maraun and Widmann, 2018) and remains an attractive alternative given the relatively small computational expense of downscaling multiple GCM simulations. Statistical downscaling has also been widely used to reduce the error in localized surface level forecasts in Numerical Weather Prediction (NWP) models. At the same time, rapid advances in machine learning, in particular deep learning, have fundamentally transformed the field of computer vision and natural language processing, with a growing

* Corresponding author.
E-mail address: peter.gibson@niwa.co.nz (P.B. Gibson).

https://doi.org/10.1016/j.wace.2022.100525
Received 7 February 2022; Received in revised form 31 October 2022; Accepted 1 November 2022
Available online 8 November 2022
2212-0947/© 2022 Published by Elsevier B.V. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

expansion into the physical sciences with promising results (LeCun et al., 2015; Silver et al., 2017; Jumper et al., 2021). In the geosciences, deep learning, in particular convolutional neural networks (CNNs), are receiving growing attention (Reichstein et al., 2019; Ham et al., 2019). A highly relevant and attractive feature of CNNs is the ability to automatically and objectively extract complex spatial-temporal information to form the basis of object detection/classification (Davenport and Diffenbaugh, 2021; Prabhat et al., 2021; Rampal and Davies, 2020) and state-based prediction (Ham et al., 2019; Chapman et al., 2019), including in the context of extreme events. This end-to-end modelling framework removes the more subjective, and likely sub-optimal, process of manually specifying thresholds (e.g. to define extreme events) or predictor features in feature engineering (e.g. through principal component analysis).

For climate downscaling, a number of studies have explored the utility of machine learning to improve upon simple statistical approaches (e.g. linear regression and analogue methods), often reporting mixed results in terms of the potential added value (e.g. Wilby et al., 1998; Schoof and Pryor, 2001; Chen et al., 2010; He et al., 2016a; Yang et al., 2018; Vandal et al., 2019). However, most of these comparison studies have used more traditional machine learning models that still require substantial feature engineering of predictor fields. Of the more recent deep learning-based downscaling approaches, single image super-resolution CNNs were first explored in detail by Vandal et al. (2017) and have since showed promising results when applied to downscaling renewable resource projections (Stengel et al., 2020). Super-resolution CNNs learn mappings between physically consistent pairs of high- and low-resolution images, which generalize across images in the series. Vandal et al. (2017) reported improved performance in downscaling daily precipitation from this approach relative to top-performing traditional statistical methods, and that this improvement increases for extreme events.

While super-resolution CNNs use a single pair of predictor-predictand fields, other CNN applications to climate downscaling have used a much wider range of predictor fields, with the expectation that these should contain additional useful information. Baño-Medina et al. (2020) tested a range of CNN architectures for downscaling precipitation and temperature over Europe finding consistent improvements over traditional methods, but also that relatively simple CNNs can often achieve superior results depending on the predictand of interest. Baño-Medina et al. (2020) highlighted that CNN performance is stable across domain sizes which can foster much larger domains than traditional methods. In a follow-up study, Baño-Medina et al. (2021) extended their reanalysis-trained CNNs to downscaling future projections from GCMs. They highlighted positive results for CNNs preserving the climate change signal from the host GCM; in contrast traditional methods may induce spurious precipitation changes to the GCM signal. Other recent studies have reported improvements for downscaling over Europe with CNNs including for temperature and precipitation (Vaughan et al., 2021) and wind extremes (Höhlein et al., 2020).

Although deep learning removes the need for manual feature selection, several important decisions remain when applying CNNs to the problem of climate downscaling. How much training data, and at what resolution, is required to learn the relevant features and non-linear relationships? What level of CNN architecture complexity is useful before overfitting? How can loss functions be implemented to target the inclusion of extreme rainfall events? What predictor variables are most important, and over what spatial domains? In this study, each of these research questions will be addressed systematically through the development and application of CNNs for downscaling mean and extreme rainfall over the New Zealand region. We also apply interpretable ML approaches to scrutinize the predictions from CNNs. For example, if a more complex CNN architecture can achieve better accuracy, what has the network specifically learnt about the relevant physical processes to enable this?

The application of CNNs for statistical downscaling is novel for the New Zealand region, where until now relatively simple linear approaches have been relied upon (e.g. Renwick et al., 2009). The New Zealand region presents a uniquely challenging region for climate downscaling, making it an ideal testbed for developing new methods. The island nation spans the southern hemisphere mid-latitude storm track and is comprised of complex mountainous terrain. Marine air masses reach New Zealand relatively uninterrupted from both polar and subtropical sources while orographic uplift in western regions can result in some of the largest annual rainfall totals on Earth (Ibbitt et al., 2000). A diversity of frontal systems and atmospheric rivers, embedded in the prevailing westerly flow, regularly drive high rainfall events across the country (Tait and Fitzharris, 1998; Prince et al., 2021; Reid et al., 2021), while extratropical cyclones occasionally reach New Zealand with frequency peaking in late summer (Sinclair, 2002). To be successful, climate downscaling must capture the full range of these processes and their relevant interactions with complex terrain at high spatial resolution.

2. Methods

The perfect prognosis approach to downscaling is implemented here (Renwick et al., 2009; Maraun and Widmann, 2018). This involves deriving empirical relationships between reanalysis (e.g. circulation fields, temperature, humidity) and surface weather parameters of interest at high spatial resolution. The main advantage of perfect prognosis is that the trained model is directly portable to other driving GCMs. This relies on the perfect prognosis assumption that the statistical characteristics and relationships derived from reanalysis fields in the training phase are a good representation of those out-of-sample (i.e. provided by other driving GCMs). Modern reanalysis data provide a suitable proxy for both NWP and GCM output. In future work, we intend to apply the trained model to both downscaling numerical weather prediction (NWP) output for short-term forecasting and downscaling GCM output for regional climate projections.

2.1. Data

Daily gridded accumulated rainfall data from the Virtual Climate Station Network (VCSN) was used here as the predictand rainfall in statistical downscaling, and as the ground truth in the out-of-sample testing period. VCSN data covers the New Zealand region on a 0.05° (~5-km) grid, derived from surface interpolation of station weather data (Tait et al. 2006, 2012). As documented for any gridded rainfall product (e.g. Gibson et al., 2019), it is important to note that precipitation biases are likely to occur in regions where station density is particularly low, namely across rugged and remote terrain like the Southern Alps (Tait et al., 2012).

Predictor fields used in model training were from the ERA5 global reanalysis (Hersbach et al., 2020). These predictor variables, detailed in Table 1, were regridded to a ~50 km grid spanning New Zealand at the

Table 1
Input training variables for the deep learning models tested for predicting accumulated daily rainfall. Training period is 1980–2012; Validation period is 2013–2016; Testing period is 2017–2020. All variables are at daily temporal resolution and 0.5° spatial resolution from the ERA5 reanalysis.

Variable name	Short name	Vertical levels	Units
Convective available potential energy	CAPE	–	J/kg
Mean Sea level pressure	MSLP	Mean sea level	Pa
Specific humidity	Q	850, 500 hPa	g/kg
Air Temperature	T	2m, 850 hPa	K
Zonal wind	U	850 hPa	m/s
Meridional wind	V	850 hPa	m/s
Vertical wind	W	850 hPa	m/s
Geopotential height	Z	850 hPa	m
Total precipitation	TP	surface	mm/day

daily temporal resolution. These variables represent both dynamical and thermodynamical drivers of rainfall across the New Zealand region. Each predictor variable was normalized relative to the mean along the spatial and temporal dimensions, and the spatially averaged standard deviation across the temporal dimension – as implemented in Rasp et al. (2020). When testing different configurations of CNNs, we assessed the importance of using different combinations of predictor variables from those shown in Table 1. For the deep learning models, interpretable ML methods (described below in Section 2.4) were implemented to explore in more detail how a particular variable was seen by the network. The input domain used in our experiments was determined through a domain of influence test. The input domain spanned 25 °S to 60 °S, and 150 °E to 170 °W. The input domain was then interpolated onto a common grid with an input shape of 64 pixels in the latitudinal direction and 96 pixels in the longitudinal direction (resolution of ~0.5 °).

2.2. Baseline linear statistical models

As described in Section 1, previous statistical downscaling approaches for New Zealand have implemented a relatively simple multiple linear regression model with stepwise variable selection (Renwick et al., 2009). This approach involved predictor variables based on the leading principal components (PCs) for geopotential height and air temperature along with other localized predictor variables (see Supplementary Material Table S1). We implement this methodology from Renwick et al. (2009) as a baseline against which skill from the machine learning models is assessed. A point of difference is that Renwick et al. (2009) relied on an earlier-generation reanalysis (NCEP/NCAR) for the predictor fields and used point-based station observations as the predictand. To allow for a more seamless comparison, we updated their earlier approach with ERA5 reanalysis and VCSN gridded predictand data.

2.3. Deep learning models

The deep learning models tested here are different configurations of Convolutional Neural Networks (CNNs). Designed to work on images or spatial data, CNNs are a regularized version of fully connected (i.e. dense) neural networks. In contrast to typical neural networks that use general matrix multiplication, CNNs approach regularization by using convolutions in at least one of their layers. This is performed by applying filters that propagate images from one layer to the next, where each filter is trained to highlight an important feature of interest across the hierarchy of complex/abstract patterns in the dataset. The CNN architecture describes the total number of trainable parameters in the model (i.e. the overall complexity) and how these are arranged and pass information across the various layers of network. The training process involves determining an optimal set of weights in the network based on applying stochastic gradient descent (Kingma and Ba, 2014) for a specified loss function. The loss function is a critical choice in the implementation of neural networks, since this determines how the network should assign errors throughout training and subsequently the final weights across the network.

A range of different CNN architectures and loss functions were implemented to test the appropriate level of complexity required to predict rainfall. While some recent CNN applications used in image recognition have used extremely deep architectures (e.g. the Resnet family with over 100 layers with skip connections, He et al., 2016b), our preliminary tests found that relatively simple CNN architectures with a more traditional encoder-decoder structure tended to give superior results. We show the performance of a subset of models (see Table 2 for descriptions) to test the added value of adding/removing convolutional layers and of including linear vs non-linear activation functions.

Using simpler architectures also enabled greater flexibility for implementing non-conventional loss functions and hyperparameter tuning of models (e.g. number of convolutional layers). Each individual model was hyperparameter tuned using a random search algorithm, in which the learning rate, dropout rate, kernel size, number of convolutional layers and batch size were randomly sub-sampled from a uniform distribution. From this, we set the parameter values fixed across models with a dropout rate of 0.6, 3 convolutional layers with a kernel size of 7 by 7 pixels in each, an initial learning rate 10^{-4} with the Adam optimizer, and a batch size of 50 days. Overfitting in each experiment is controlled through dropout regularization, an adaptive learning rate and early stopping. The number of epochs required varies across different experiment and model architectures. However, typically 40–60 epochs are required for training with each model taking 2–4 min to train on a NVIDIA Tesla P100.

We compared the most-commonly used Mean Squared Error (MSE) loss function (e.g. in regression applications) alongside the Bernoulli–gamma distribution (see Table 2), based on the performance of this approach shown by Baño-Medina et al. (2020) for downscaling precipitation with CNNs, which was introduced in Cannon (2008). The log-likelihood of the Bernoulli-gamma distribution, defined hereon as Gamma loss, can be defined as below:

$$L(p,\alpha,\beta;y_{true},y_{pred}) = (1-p_{true})\log(1-p+\varepsilon) + p_{true}\,(\log(p+\varepsilon) + (\alpha-1)\log(y_{true}+\varepsilon) + \alpha\log(\beta+\varepsilon) - \log(\Gamma(\alpha+\varepsilon))) - y_{true}\,/\,(\beta+\varepsilon) \quad (1)$$

Where $L(p,\alpha,\beta;y_{true},y_{pred})$ is the Bernoulli-gamma (Gamma) loss function, $p_{true} = p_{true}(y_{true})$ is binary rainfall (e.g. 1 for rainfall, 0 for no rainfall), α and β are the shape and scale parameters of the Bernoulli-gamma distribution, p is the probability of rainfall predicted by the neural network, y_{true} is the rainfall amount, Γ defines the Gamma distribution and ε is the regularization parameter. The expected value of the distribution, namely $y_{pred}(p,\alpha,\beta)$ defines the final predicted rainfall, with the CNN tasked with predicting these parameters for a given day and grid-point (see output layer from Fig. 1). Here, both the binary and continuous nature of rainfall are captured in a single loss function.

Table 2
Details of the 5 main deep learning models evaluated. Models were hyperparameter tuned for dropout value, number of dense layers and number of convolutional layers. Further sensitivity testing of increasing model complexity was not found to significantly improve performance (either through adding additional convolutional or dense layers). The output layer for Gamma loss function consists of three neurons for each grid point to predict the parameters of the Bernoulli-gamma distribution.

Model Name	Non-Linear CNN (Gamma)	Linear CNN	Non-Linear Gamma	Linear Gamma	Linear Dense
Trainable parameters	10.75M	4.85M	16.72M	16.72M	10.82M
Regularization	Dropout	Dropout	Dropout	Dropout	Dropout
Batch Normalization	Yes	Yes	Yes	Yes	Yes
Loss Function	Bernoulli-Gamma	MSE	Bernoulli-Gamma	Bernoulli-Gamma	MSE
Dense Activation	Relu	Linear	Relu	Relu	Linear
Dense Layers	[256] -> [11491,3]	[256] -> [11491]	[256] -> [11491,3]	[256] -> [11491,3]	[256] -> [11491]
Convolutional Filters	16, 32, 64	16, 32, 64	–	–	–
Kernel Size	(7,7)	(7,7)			
Convolutional Activations	Relu	Linear	–	–	–
Pooling	Max Pooling	Max Pooling	–	–	–

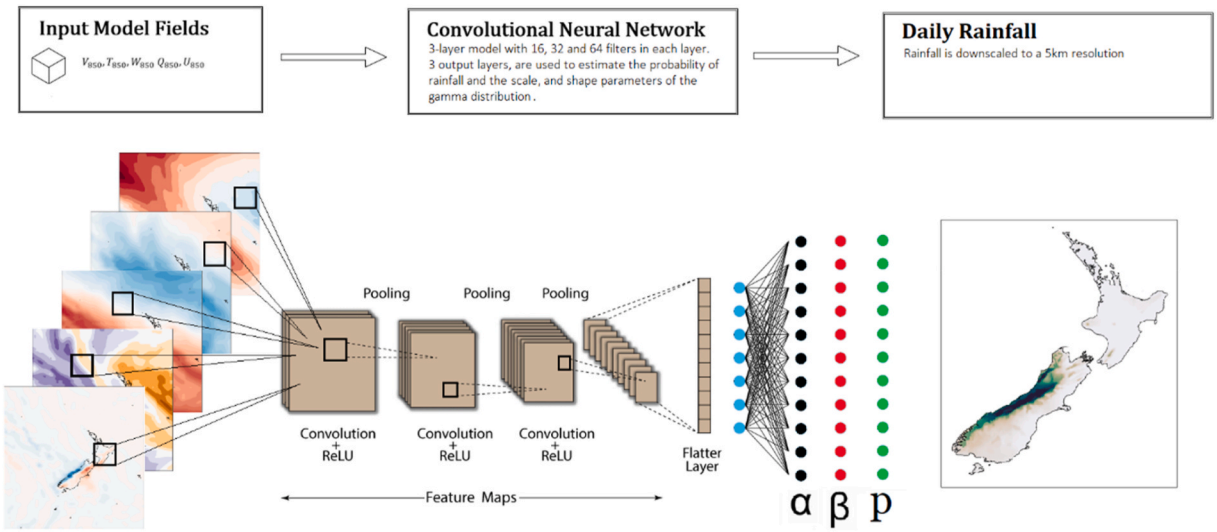

Fig. 1. Summary of the CNN approach for downscaling rainfall. A subset of the candidate input fields used as predictors in the model are shown (see also Table 1). Both predictors and predictand are at the daily temporal resolution. The CNN output consists of three layers for the probability (p) of rainfall and the shape (α) and scale (β) parameters from the gamma distribution, where the number of neurons in each output layer is equal to the number of grid cells over the domain. These parameters are then used to predict the daily accumulated rainfall at each grid cell on a 5 km gird over New Zealand.

2.4. Machine learning interpretability

Interpretable machine learning (also referred to as 'explainable AI') addresses a longstanding criticism that deep learning models are 'black boxes' due to the perceived inability to understand how a prediction was made. To address this, in the context of deep learning, interpretable machine learning methods attempt to visualize how certain locations or predictor variables influence a decision/prediction made by the network. A model that is both accurate and interpretable has been emphasized as important for building trust in the prediction (Toms et al., 2020; Gibson et al., 2021) or may even lead to the development of new understanding in complex systems (e.g. Mayer and Barnes, 2021). Here we implement Gradient-weighted class-activation maps (Grad-CAM, Selvaraju et al., 2017) to better understand how the CNN made a certain rainfall prediction for a given location. In particular, Grad-CAM identifies which spatial locations in the input fields, for a given depth in the network, most strongly support the prediction from the network at a given output location (e.g. rainfall over a given city). Grad-CAM does not specify which individual predictor variable/channel in the network had the strongest influence, rather it shows spatial locations in the aggregated predictor space that were most important. To better aid interpretation, for our overall best performing CNN, we subsequently trained different CNNs (keeping the architecture fixed) with different subsets of predictor variables to visualize their contributions to the network. To limit the number of results shown, we focus on a network trained with predictor fields MSLP, Q850, T850.

In similar weather and climate applications, saliency maps (McGovern et al., 2019; Rasp and Thuerey, 2021) and layerwise relevance propagation (LRP, Toms et al., 2020; Barnes et al., 2020; Ebert-Uphoff and Hilburn, 2020; Mayer and Barnes, 2021) have also been implemented, but both carry some disadvantages depending on the specific application. Saliency maps do not necessarily imply importance, rather they show how the output changes when a set of predictor values are slightly perturbed. Saliency maps also tend to focus on local gradients in the input space whereas a more global view is often required (McGovern et al., 2019; Ebert-Uphoff and Hilburn, 2020). LRP has many attractive features for weather and climate applications (e.g. Toms et al., 2020), however a disadvantage is that LRP is only available for a relatively small set of neural network architectures, whereas Grad-CAM can be applied more widely (Selvaraju et al., 2017).

3. Results and discussion

3.1. Performance relative to existing models

We begin by discussing performance of the top-performing CNN relative to an existing and commonly used statistical downscaling model based on Renwick et al. (2009) which implements multiple linear regression (Fig. 2). It is important to note that all comparisons are made on the 'test set' data (years 2017–2020) which were unseen by either model in the training and verification phases. The validation set (years 2013–2016) is used for hyperparameter tuning and evaluation of the model performance during training. In the next subsection, we will compare this top-performing CNN against other candidate deep learning models (Table 2) to shed light on what gives rise to improved performance. As shown in Fig. 2a, in terms of the temporal variability on wet days, the CNN has much improved performance with $R^2 > 0.5$ across much of the country. In contrast, the linear model's performance varies considerably by region, with degraded performance across eastern parts of the South Island (Fig. 2b). The median improvement in explained variance (R^2) is 0.17, with the spatial distribution of the improvement shown in Fig. 2c. The root-mean-squared error in precipitation is also reduced in the CNN model, with a reduction in percentage error by over 20% (Fig. 2f). Rainfall extremes are also better resolved in the CNN model in most regions, the 90th percentile rainfall amounts are dry biased by >40% in the linear model, with this dry bias much reduced by over 25% nationally (absolute change) in the CNN model (Fig. 2i).

There is still a general tendency for larger errors in some high elevation locations across the country for both models. This is likely due to the higher total rainfall in these regions, however the higher observational uncertainty in these regions may also contribute through creating systematic biases in the model training. Simply adding elevation as an additional predictor variable was not found to alleviate this bias (not shown). Despite the improved performance, it is interesting to note that CNN model still has an overall dry bias for extreme rainfall. This dry bias in extreme precipitation was also reported in the studies of Baño-Medina et al. (2020) and Baño-Medina et al. (2021) who applied similar CNN architectures for downscaling over Europe, and in Sun and Lan (2021) for downscaling over China.

In terms of their relative computational costs, perhaps counterintuitively, the more complex CNN model is faster in both the training and prediction phases. This is because the standard linear modelling

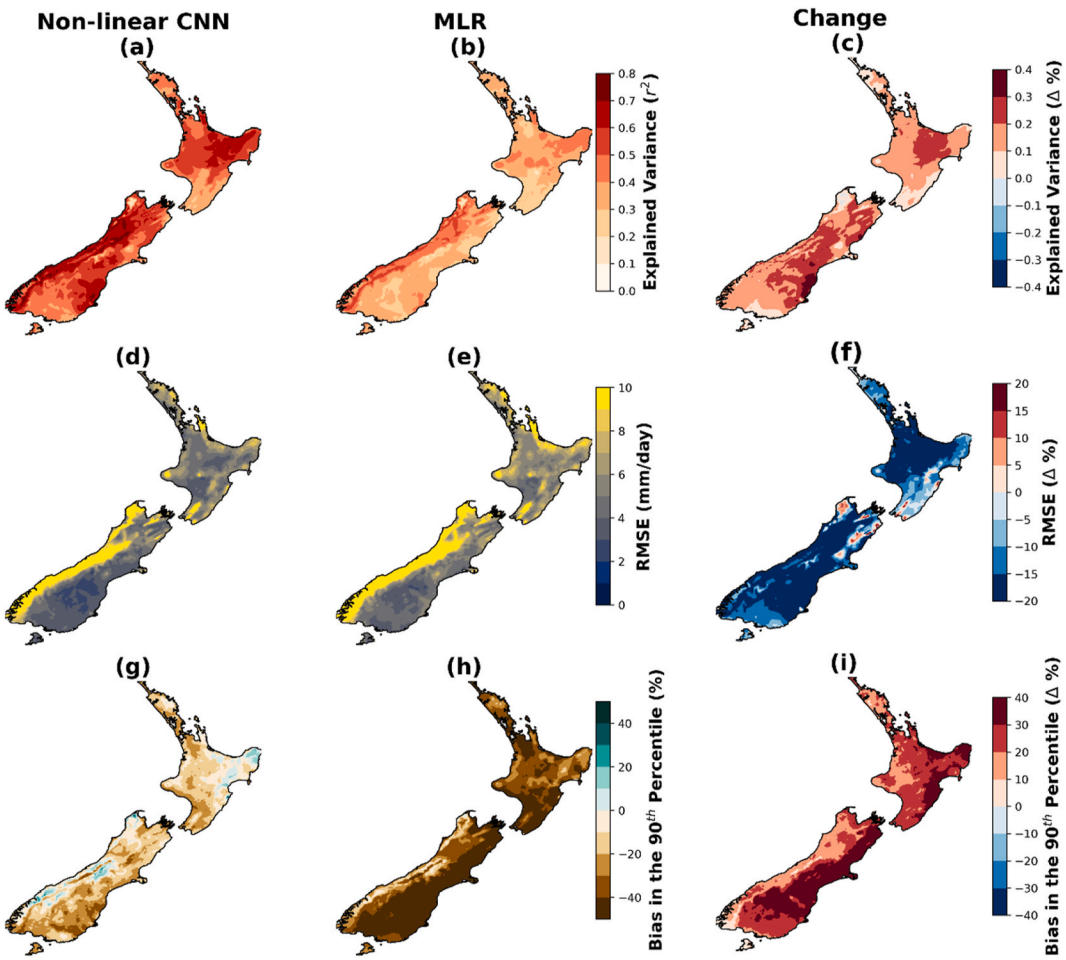

Fig. 2. Comparison of the top-performing CNN model (non-linear CNN with Bernoulli-gamma loss function) against the multiple linear regression model (MLR) from Renwick et al. (2009) in terms of: Explained variance on rainfall days (R^2), root-mean-squared error (RMSE) on rainfall days, and percentage bias of the 90th percentile of rainfall. The third column outlines the difference (%) between the CNN and MLR. For (c) and (i) the difference between the top performing CNN and MLR are computed as the absolute difference (CNN – MLR), and for (f) the difference is computed as the relative percent difference.

approach requires training a separate model for each predictand grid point separately, whereas the CNN model performs training simultaneously for all grid cells (i.e. across the output layer shown in Fig. 1). In an operational context, the compute time for implementing thousands of linear models to statistically downscale NWP or GCM output in parallel would be several orders of magnitude larger than downscaling through a single CNN. To highlight the computational improvements in both training time and prediction time, we benchmarked the non-linear CNN model against the parallelized MLR model (Supplementary Material Table S2). To statistically downscale rainfall at each individual VCSN grid point requires a total of 11491 regression models. While any individual MLR model is less computational expensive and complex than the non-linear CNN, the magnitude of MLR models required means that it is overall less computational efficient than a single CNN. We conservatively estimate a speed up of at least 150 times through using an individual CNN. While further optimization and parallelization of MLR is possible, it is clear that non-linear CNNs are computationally cheaper.

3.2. Deep learning architecture considerations

We now discuss the neural network architecture which has led to the documented increase in performance. To do so, we train and compare performance from five candidate neural network models based on systematically adding and removing convolutional layers, non-linear versus linear activations, and different loss functions (Table 2). As shown in Fig. 3, the non-linear CNN with Gamma loss functions typically performs the best, in terms of explaining temporal variability (Fig. 3a) and for reducing root-mean-squared error (Fig. 3b) and extreme rainfall error - defined as the percentage bias the in the 90th percentile of rainfall (Fig. 3c). On the other hand, the dense neural network, with linear activation functions, no convolutional layers and the simpler MSE loss function, shows substantially degraded performance. Comparing across models, the largest individual improvement appears to stem from switching from the MSE loss function to the Gamma loss function. The advantages from using the Gamma loss are further illustrated in Fig. 4 when comparing downscaled rainfall across rainfall percentiles. The added value of the Gamma loss function is considerably larger for capturing more extreme amounts of rainfall. This is likely because the Bernoulli-gamma distribution incorporates both the binary and continuous nature of rainfall and is well-suited for capturing spatial relationships in rainfall between locations (Cannon, 2008). Improvements are also shown when using a CNN with MSE loss versus the linear dense model with MSE loss. This suggests that adding convolutional layers with CNNs aids in the learning of complex spatial relationships from large-scale fields, which traditional dimensionality reduction methods fail to capture.

3.3. Input data considerations

Another key consideration involves deciding which input data to present to the CNN during training. This includes selection of predictor variables, domain size, and the number of samples. Training multiple

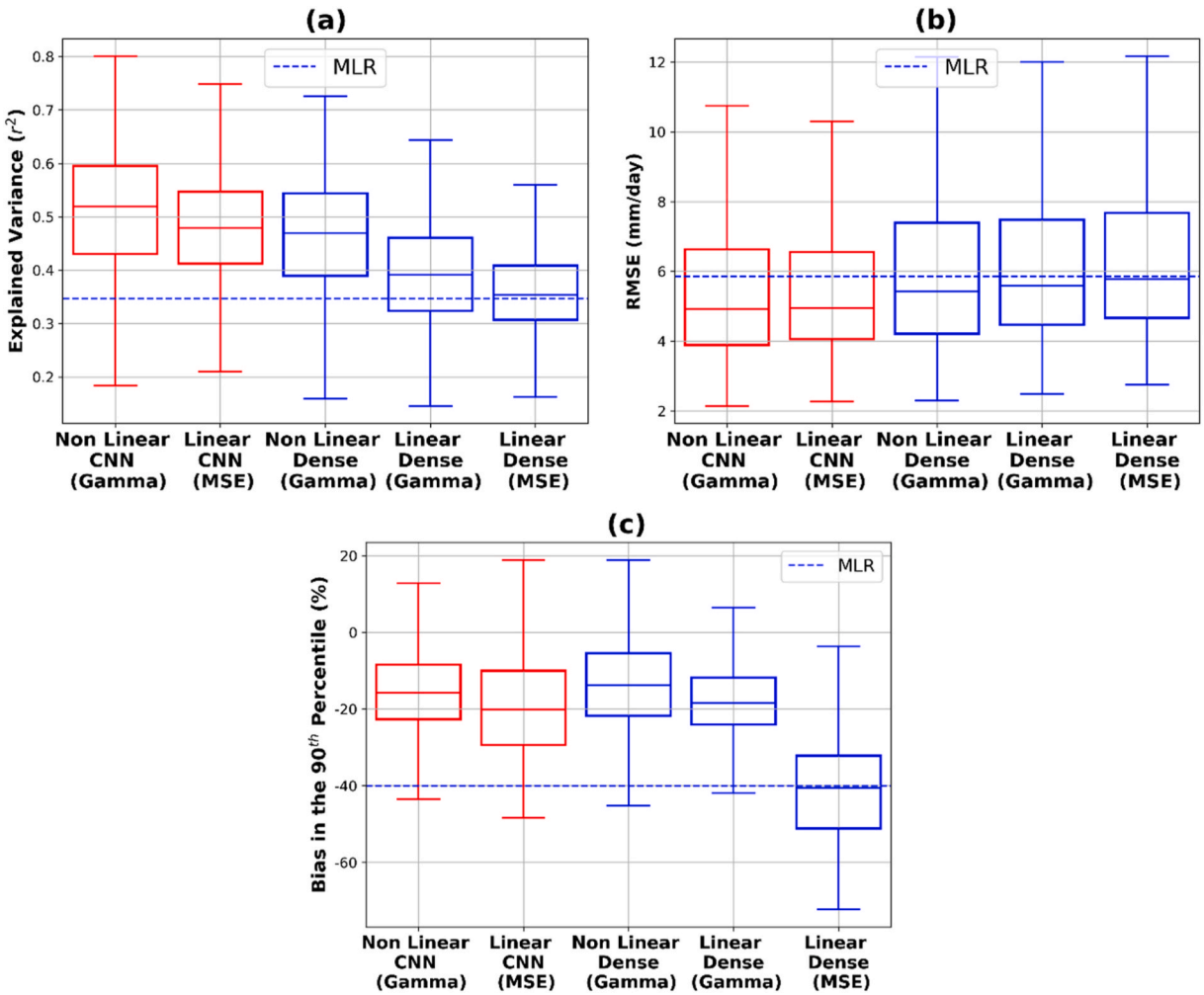

Fig. 3. Comparison of network architecture and loss functions in terms of downscaling performance. Downscaling performance is assessed from three metrics, explained variance (a), root-mean-squared error (b) and percentage bias in the 90th percentile of rainfall (c). CNN models are shown in red and dense NN models are shown in blue. For a given model/performance indicator, the spread relates to the variability across all grid-cells. The Non-Linear CNN (Gamma), which is generally the top-performing model, was shown earlier in Fig. 2.

CNNs with different subsets of the full predictor variable set (Table 1) illustrates which combinations of predictor variables were overall most important (Fig. 5). It is well established in applied machine learning that continually adding additional information does not necessarily translate to an overall improvement in accuracy on the test data. This is primarily due to the propagation of redundant information under strong correlations between predictor variables, which can be common across meteorological variables. We focus predominantly on downscaling atmospheric fields at the 850-hPa level as preliminary tests found this to generally be the strongest predictor of rainfall (compared to 700-hPa and 500-hPa levels).

Overall, three of the most important predictor variables are shown to be the wind components (U, V, W) at 850-hPa. While these perform well in combination, individually the contributions from U and V are small in terms of explained variance. This is expected given that the orientation of the wind field is an important driver of rainfall which can only be deduced from combining these variables. When combined with the wind components, additional variables including temperature (T850), humidity (Q850) and total precipitation (TP) result in additional improvements in performance. These predictor variables have been commonly applied in various statistical downscaling models, as detailed in the large intercomparison project of Gutiérrez et al. (2019). Interestingly, using total precipitation from the driving reanalysis as a single predictor variable still results in reasonably good performance. Such an approach is equivalent to the "super-resolution" application of CNNs (e.g. Vandal et al., 2017; Stengel et al., 2020). Our results suggest that there is significant value going beyond a "super resolution" type approach, and the combination of other prognostic variables can add value to downscaling rainfall. It is encouraging that the CNN performance shown here, as a function of the different combinations of predictors, aligns well with existing physical understanding of regional rainfall drivers. For example, individual predictor variables such as temperature are shown to perform poorly since this does not directly include information regarding either the larger-scale circulation or moisture availability (Gutiérrez et al., 2019).

While only select variable combinations are shown in Fig. 5, we find negligible improvements in skill from adding more than five predictors. The additional atmospheric variables at other levels in the atmosphere (e.g. 500-hPa) also did not show any significant improvement over a five-variable combination. While further investigation of feature importance at different atmospheric levels is warranted, our results have indicated that contributions from other atmospheric levels are not required. It is likely that information from these upper-level fields (e.g. jet variability) is indirectly captured in the lower-level atmospheric fields to a sufficient degree. We also note that different applications of downscaling may require different variable considerations, as discussed in Gutiérrez et al. (2019). For example, if downscaling rainfall for future climate change projections, it would become important to include

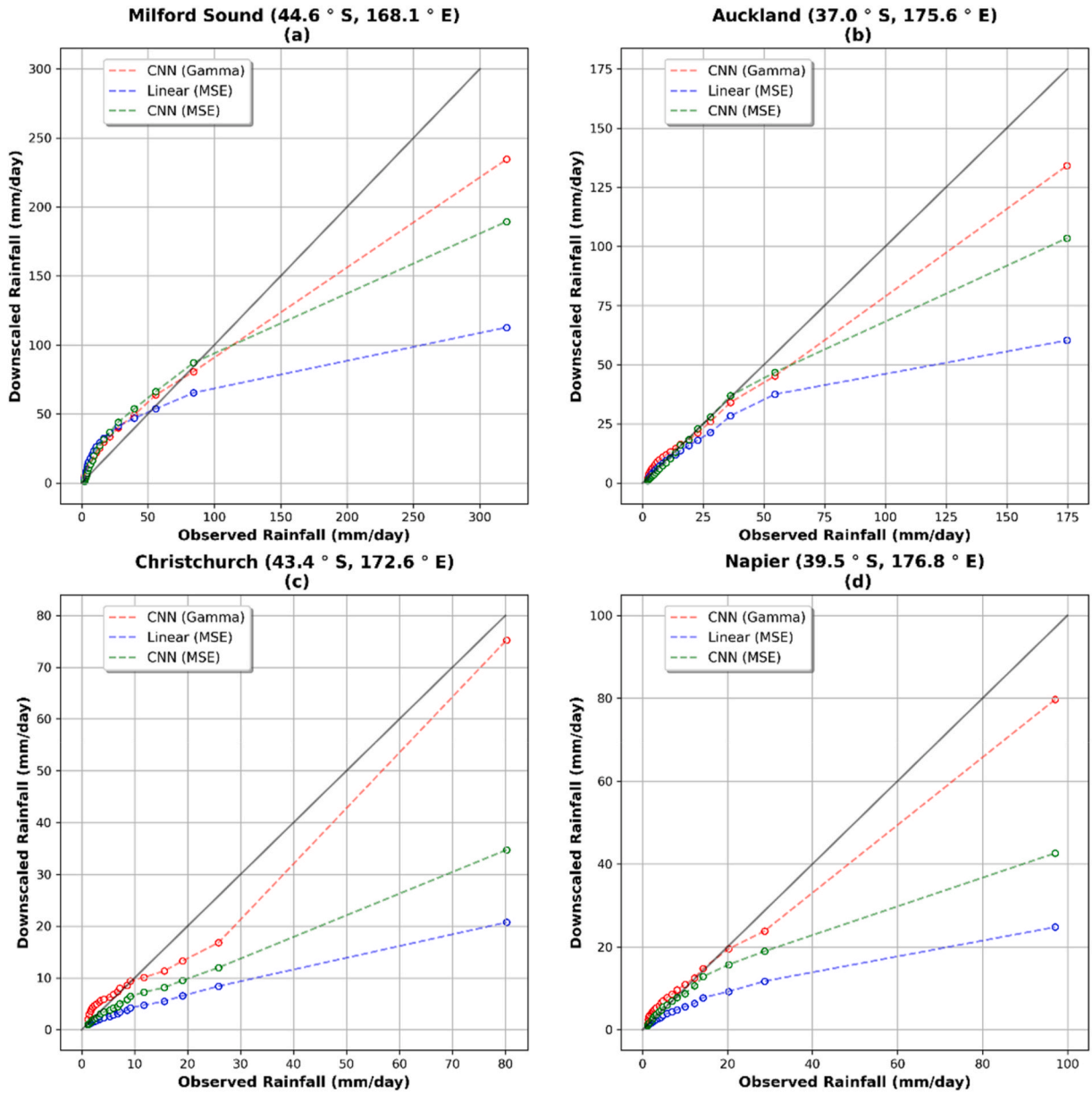

Fig. 4. Q-Q plots showing the relationship between observed and downscaled rainfall percentiles at four locations which span New Zealand: Milford Sound (a), Auckland (b), Christchurch (c) and Napier (d). Values closer to the black y = x line indicate better performance. The percentiles are computed only for only rainfall days (>1 mm), computed every 5th percentile out to the 99th percentile. As shown, the CNN with gamma loss function performs considerably better at resolving extreme events.

thermodynamic-related predictors that are non-stationary with climate change (e.g. specific humidity and temperature).

The downscaling performance was also explored in terms of varying the input domain size (Fig. 6) and number of training samples (Fig. 7). We identified two important predictand locations (Auckland and Wellington, based on population density) to separately downscale daily fields of U, V, W, T and Q. The domain of influence outlined in Fig. 6 is defined as the predictor field domain in length around these locations. This experiment probes the utility of various architectures in capturing information from large-scale atmospheric flows. The downscaling performance is shown to generally increase with increasing domain size but becomes asymptotic after a certain point. For the CNN, the added value from increasing domain size is typically found to asymptote after ~14° in the input spatial domain, after which both the error and explained variance converge (Fig. 6). This indicates that spatial features (e.g. gradients) at scales greater than ~14° would not likely add further value. This also highlights that CNNs are capable of learning non-local influences more readily than traditional linear approaches, and that learning spatial features and gradients over a larger domain (e.g. up to ~14 °) may help condition local-scale variability. The difference in explained variability between the linear dense and the CNN as a function of domain size is notable here. For small domain sizes (<2 °), the difference in explain variance is approximately 5% for both locations. This difference grows to approximately 10% for domain size larger than 10 °, suggesting that the CNN is better suited to extracting more complex non-linear relationships over the larger spatial domain which is not deemed useful information by the simpler linear model.

The relationship between model performance and training data size is shown in Fig. 7. As expected, this relationship is roughly logarithmic, with a rapid increase in the proportion of explained variance for small datasets (1–10 years) and slower growth in performance when the data size increases beyond 10 years. Notably, the proportion of explained

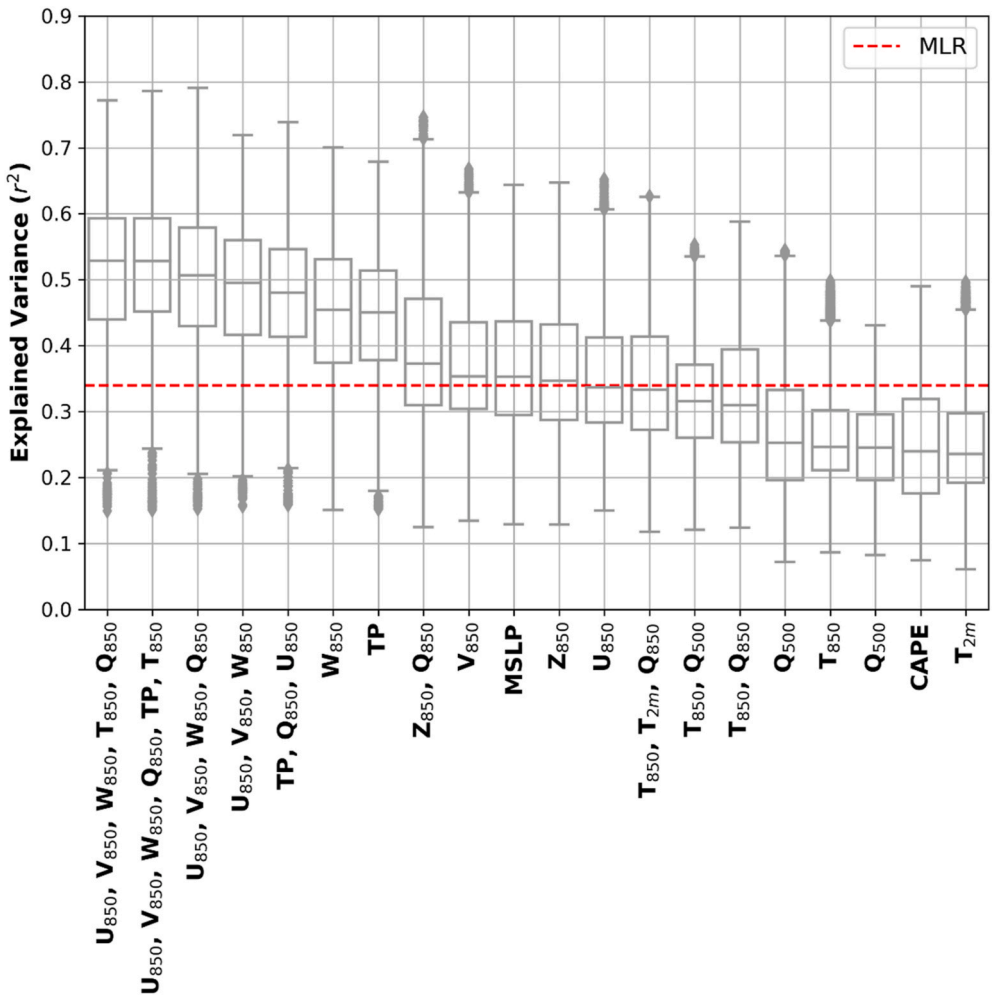

Fig. 5. Explained variance in rainfall (>1 mm) from various combinations of input fields in the CNN model. The top performing models select combinations of wind speeds (zonal, meridional and vertical: U,V,W), specific humidity (Q), temperature (T) and total rainfall (TP). These combinations were selected from the candidate variables and levels in Table 1.

variance with a linear model after 35 years of training data is exceeded by the CNN model with just five years of training data. The rate at which the amount of variance explained increases is also much lower for the linear dense model compared to the CNN. As discussed earlier for domain size considerations (Fig. 6), this suggests that the CNN is benefiting more readily from the larger number of input cases. Furthermore, Fig. 7 illustrates that even with 35 years of training data (>12,000 daily input fields) the performance for the CNN is yet to converge. As such, it is plausible that additional training data could further benefit downscaling performance. Using extended historical observational records (e.g. in the pre-satellite era) is one potential path forward, though this would likely introduce additional uncertainty associated with these earlier records. A different approach, described in more detail below, would be to implement transfer learning where the network is pretrained on long model simulations to provide multiple independent realizations of the system. In the final stage, the network weights are then adjusted over the shorter historical observational period which also adds a degree of bias correction (e.g. Ham et al., 2019).

The particular training, validation, and test set years were primarily chosen to maximise the amount of training data available in the learning process, which was shown earlier to be important (Fig. 7). However, it is possible that the relatively short test set period (4 years, 2017–2020) may not be sufficiently long to provide a robust characterization of model performance for extreme events. To examine the sensitivity of this, we repeated our analysis (Supplementary Material Figs. S1–S2) by extending the testing period to 10-years (years 2011–2020) with a slightly reduced training (years 1980–2009) and validation period (year 2010). As shown in Supplementary Material Fig. S1, the improvements of the non-linear CNN over the MLR approach remain consistent for this longer testing period, including for extreme events. There are some slight spatial differences in the skill between test set years, which is to be expected given natural variability across years and the differences in the length of the training period. The robustness for extremes is also shown in the Q-Q plots across this longer test period (Supplementary Material Fig. S2) where the CNN (Gamma) typically shows superior performance. Sensitivity tests also revealed that a relatively short 1-year validation period was sufficient in this analysis. It is important to note that the validation dataset is used mainly to assess overfitting, in which case the learning rate hyperparameter is either reduced (learning rate scheduling) or training is stopped (early stopping).

3.4. Interpretable deep learning

Using Grad-CAM, we next scrutinize where the network is 'looking' when making predictions of extreme rainfall events at select locations (Fig. 8). To do so, we present two case studies showcasing different types of extreme rainfall events. To illustrate where the network is 'looking', we have computed the normalized importance for each grid point (Fig. 8a and d) as contributing to the model's prediction of daily rainfall

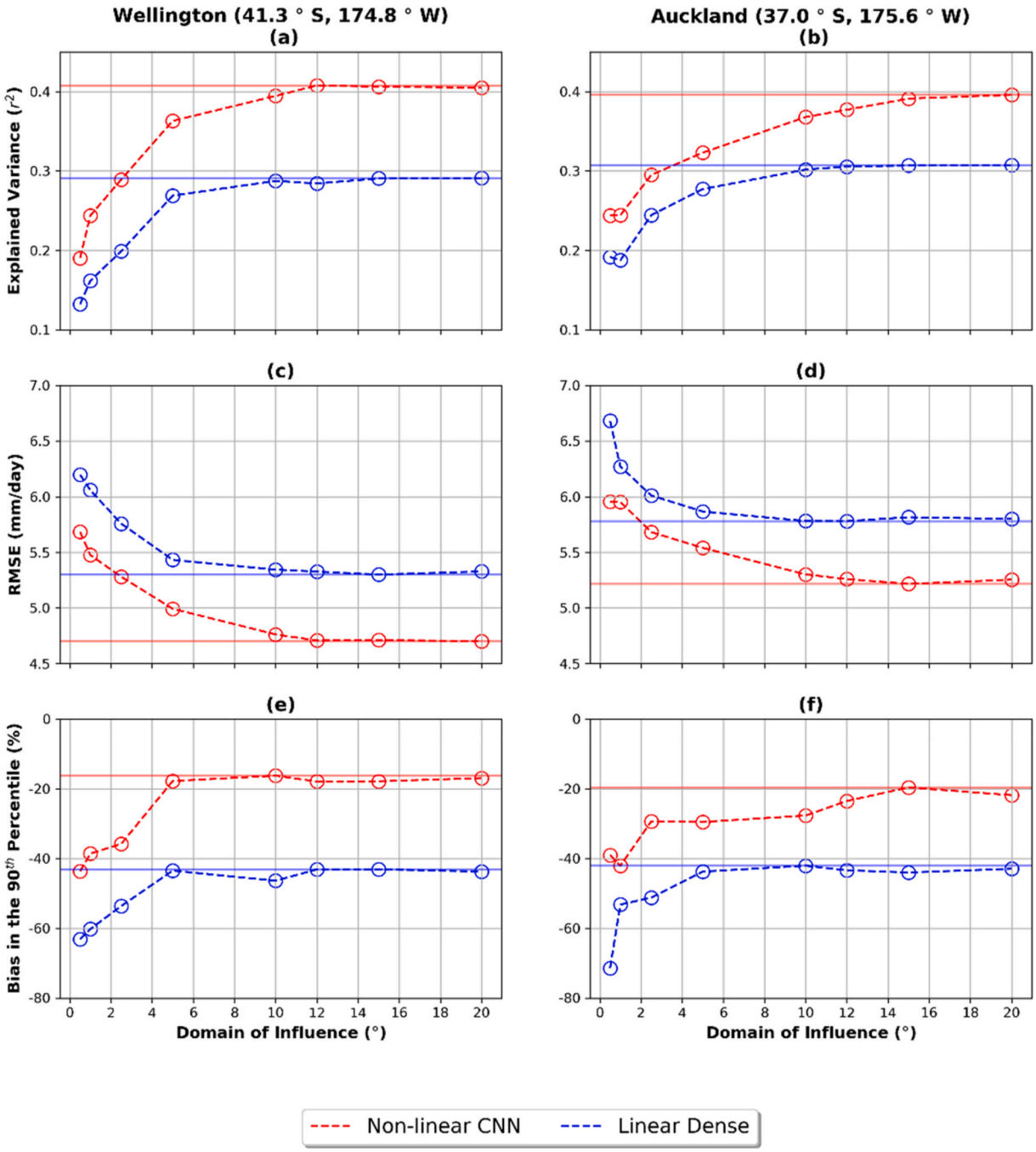

Fig. 6. Explained variance (a,b), root-mean-squared error (c,d) and the percentage bias in the 90th percentile (e,f) in rainfall as a function of domain size (° longitude and latitude) of input fields for two select locations, Wellington and Auckland. The solid horizontal line refers to the minimum error for a given model.

(Fig. 8c and f). The normalized importance is computed across MSLP, Q850 and T850 variables combined to highlight regions of combined importance across these variables. The observed background circulation and moisture fields for the events are also shown (Fig. 8b,c,e,f).

We first probe an extreme atmospheric river event that caused severe flood damage along the west coast of the South Island and broke rainfall records (Prince et al., 2021), shown in Fig. 8(a–c). Here, the network is shown to focus on the footprint of the atmospheric river, including the band of tropical vapor transport from the north-west and the extended region of extreme vapor transport over the south-east of the country in the lee of the event (Prince et al., 2021). This further illustrates the ability of the model to learn both 'local' and 'non-local' information. Note, in this instance the networks prediction resolves the spatial distribution of rainfall well for the event, where the prediction of rainfall (Fig. 8c) closely matches the observed rainfall (Fig. 8b).

The second example was driven by a deep ex-tropical cyclone (named Cyclone Gita) that passed over the northern region of the South Island following a sharp north-westerly track, as shown in Fig. 8(d–f). The cyclone caused substantial damage across several Pacific Island nations before reaching New Zealand where it was responsible for widespread flood and wind damage. Here we target rainfall over Kaikōura since this was a location of extreme rainfall in both observations and the downscaled (i.e. predicted) rainfall. As shown in the normalized importance heatmap (Fig. 8d), the CNN was focusing attention on the bands surrounding the centre of the cyclone. A more expansive region to the north of the event was also highlighted by the

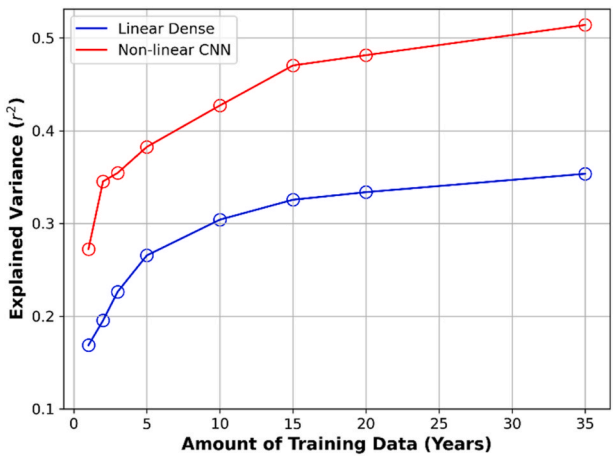

Fig. 7. Fraction of explained variance in rainfall as a function of training data size available to the network.

network, likely due to the extended north-westerly flow and elevated atmospheric moisture content (Fig. 8e and f).

While these two events represent only a small fraction of extreme rainfall events and their associated meteorological drivers, it is encouraging that in both cases the network has been capable of targeting the primary meteorological features of importance, including local and non-local information. Notably, both these cases are derived from the test dataset, where the CNN did not have access to the predictand rainfall information, indicating that these relationships can generalize well outside of the training dataset. We also note that in these two examples, locations that did not receive rainfall had a considerably different importance heatmap and did not regard the atmospheric river or cyclone important to the predicted rainfall, as would be expected (not shown). These results serve as a useful test for building credibility and trust in the predictions from deep learning models that are otherwise difficult to interpret. In contrast, Lapuschkin et al. (2019) present examples from the field of image classification where CNNs can achieve high predictive accuracy for the wrong reasons, i.e. learning spurious correlations within large datasets that carry no plausible physical interpretation.

3.5. Related work and future directions

Given the improvements presented here with deep learning, it is reasonable to question how close we might be to a performance ceiling for downscaling rainfall, based on imperfect input information from the large scale. Here we discuss some potential avenues for future research. Our current CNN implementation does not capture temporal relationships in the data; instead, each concurrent input field is matched to a single output field. However, temporal autocorrelation is well known to exist in meteorological variables, which for the relevant synoptic variability is on timescales of a few days to weeks. As such, adding long short-term memory (LSTM) layers to the network (i.e. recurrent neural networks) could capture some of this information to potentially improve performance (e.g. Hayatbini et al., 2021). For example, capturing the way in which a storm migrates into a given region (e.g. speed, orientation) in the lead up to a rainfall event may provide additional predictive information beyond that of the concurrent information alone. In this implementation, the inclusion of sub-daily predictor variables may

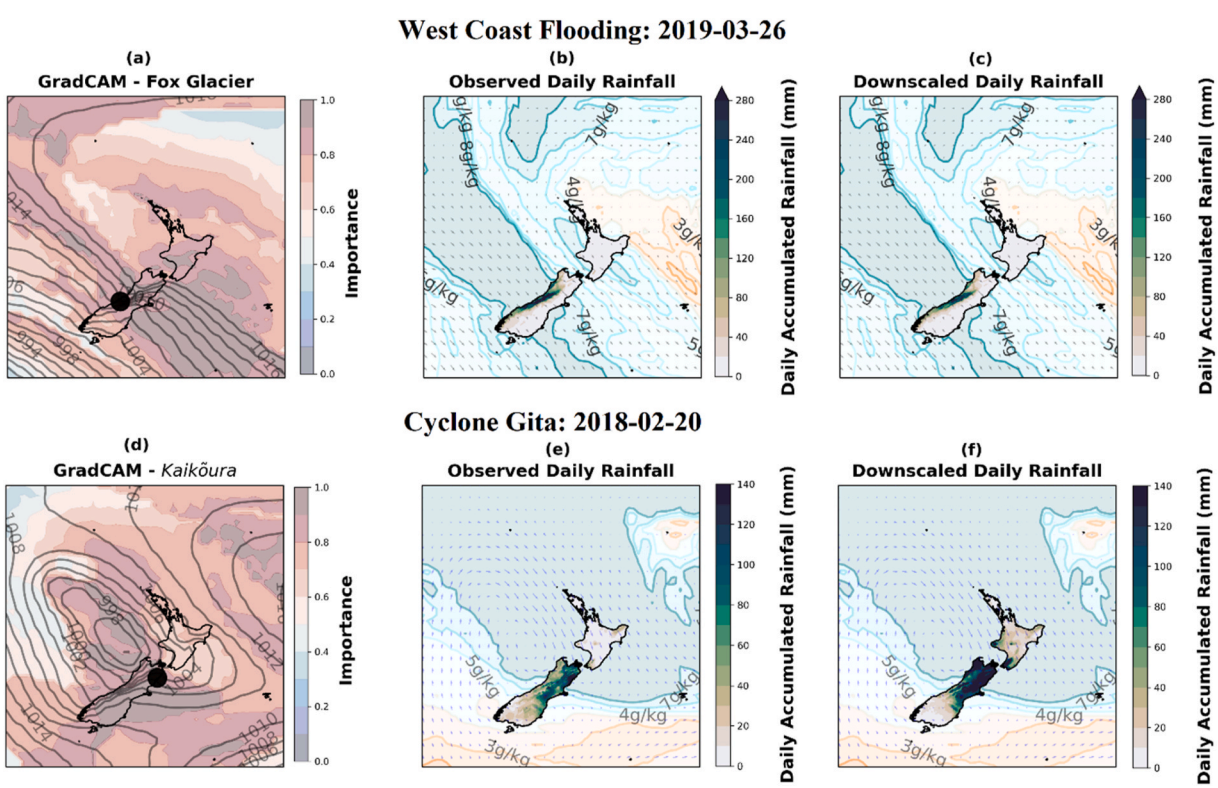

Fig. 8. Implementation of Grad-CAM to explore what the CNN has focused on in the input fields when downscaling rainfall. Two extreme events were examined: a West Coast Flooding event (panels a–c) and Cyclone Gita (panels d–f). Panels a and d highlight the normalized importance from Grad-CAM at each grid point in the input data for the model's prediction of rainfall at Fox Glacier (black circle, panel a) and Kaikōura (black circle, panel d). When aggregated across all predictor fields, the area highlighted in red is most important to the model's prediction, with observed MSLP contours also overlaid. Panels b and e show the observed daily accumulated rainfall for the individual events, with the wind fields (vectors at 850-hPa) and specific humidity fields (shading) overlaid. Panels c and f show the model's prediction of daily accumulated rainfall for the individual events, with the same observed wind fields and specific humidity fields overlaid. Here the network is trained with predictor fields: MSLP, Q850 and T850.

become more important to better account for these finer migratory details of the storm moving into the region.

While more complex and varied neural network architectures were examined here, these were far from exhaustive. A rather different (i.e. unsupervised) type of architecture is based on conditional generative adversarial networks (cGANs, Goodfellow et al., 2014; Mirza and Osindero, 2014). This approach, when tailored to the problem of forecasting rainfall, has shown promising results recently (Hayatbini et al., 2019; Ravuri et al., 2021). The underlying approach of cGANs involves two 'adversarial' models: a generator network which is responsible for generating images and a discriminator network tasked with evaluating the output from the generator. In applying cGANs to high resolution precipitation nowcasting, Ravuri et al. (2021) used two loss functions in the training phase: a separate spatial and temporal discriminator. The spatial discriminator is a CNN that determines whether fields are degraded or not to remove certain undesirable outcomes (e.g. spatially smoothed predicted field). The temporal discriminator imposes temporal consistency in temporal sequences to remove physically inconsistent or jumpy sequences of predictions. These discriminators, combined with additional regularization, were shown to achieve a high degree of spatiotemporal consistency and outperform several candidate deep learning models. To the best of our knowledge, a similar approach has not yet been explored in the context of climate downscaling.

As mentioned earlier, transfer learning is another plausible direction towards improving downscaling performance. A fundamental challenge of obtaining accurate and robust results from deep learning relates to out-of-sample generalizability, which typically improves with sample size. Indeed, evidence for this was shown in Fig. 7 where accuracy, to some extent, continues to improve with training sample size. Relative to other fields, climate science is in a fortunate position whereby the amount of 'synthetic' data simulated by climate models is extremely large and continues to grow. For climate downscaling, given the need for high spatial resolution data in the predictand during training, regional climate model simulations from coordinated experiments like CORDEX could be heavily leveraged for transfer learning.

A potential limitation of the current approach is the use of the gridded rainfall product (VCSN) as the predictand in the training phase and as ground-truth in the testing phase. Indeed, VCSN is known to have rainfall biases, particularly in high-elevation areas of the South Island where the observation network is sparse (Tait et al., 2012). While VCSN is arguably the best currently available high-resolution gridded rainfall product for New Zealand, these biases, which remain difficult to quantify, may flow into the relationships established through downscaling to degrade performance.

Lastly, we make a general comment about the ongoing need for continued investment in both dynamically-based regional climate models (e.g. CORDEX) and machine learning-based models. While current and future applications of deep learning may begin to show superior downscaling performance in certain settings, with considerably lower computational overhead, there are also important limitations that will not likely be solved in the short term. Namely, dynamical models are based on well-understood physical principals (despite still requiring substantial parameterizations) which allow for various non-linear physical feedbacks to be represented in a changing climate (i.e out-of-sample). In contrast, current ML approaches to downscaling, that have learnt only from historical data, may fail to capture certain non-stationary processes (Watson-Parris, 2021; Kashinath et al., 2021). Implementing loss functions that represent well-known physical constraints in the climate system (Kashinath et al., 2021), or conditionally training the network on future climate states from model simulations, may partly address this issue. In future work, we will comprehensively examine the ability of the trained ML approach to generalize across non-stationary historic and future climate conditions. This will include carefully designed tests that contrast hot-versus-cold years in the historical training and testing years (Salvi et al., 2016), as well as comparing the realism of the future climate change signal between ML models and dynamical models for New Zealand. These tests, while outside the scope of the present work, will be necessary before applying the trained model to future climate change projections.

4. Conclusions

For the first time, we have explored the potential for a deep learning framework to improve rainfall downscaling over New Zealand. The best performing CNN model is capable of outperforming existing statistical approaches, both in terms of temporal variability and in terms of mean and extreme rainfall. The main findings are:

- When spatially aggregated across the region, the fraction of explained variation on wet days increased from 0.35 to 0.52 (Fig. 2c). The existing dry bias for rainfall extremes decreased from approximately 40%–~15%, Figs. 2i and 3).
- The improvements from implementing deep learning are largest for extreme rainfall events (Figs. 2i and 4). In terms of network architecture, the largest benefits stem from implementing a probabilistic loss function. Improvements also come from adding convolutional layers and non-linear activations.
- Combinations of predictor variables that consistently improve performance are well-aligned with physical intuition (Fig. 5). For example, a network trained on atmospheric circulation and available moisture fields far outperforms a network trained solely on thermodynamic fields.
- Increasing the domain size over which predictor fields are sampled and increasing the number of training samples generally improves out-of-sample downscaling performance. The domain size has less bearing on performance for linear models, suggesting that the non-linear CNNs are better suited for extracting complex information across an extended domain.
- Through implementing interpretable machine learning techniques (Grad-CAM), we explored the learned relationships in the CNN. Across different cases of large-scale rainfall drivers (e.g. atmospheric rivers and ex-tropical cyclones), the trained CNN could target the most relevant meteorological features when making predictions of extreme rainfall. This suggests that the model is capable of learning complex and physically plausible relationships.

Avenues for further improving downscaling performance through deep learning were discussed. These include testing a broader range of architectures including adding LSTM layers or implementing cGANs which have shown recent promise in the fields of rainfall nowcasting. Transfer learning, whereby the CNN is pretrained on high-resolution model simulations to greatly increase the training data sample size, is another promising avenue.

Author statement

Neelesh Rampal: Conceptualization, Methodology, Data Analysis, Visualization; **Peter Gibson:** Conceptualization, Methodology, Writing, Supervision, Reviewing, Editing; **Abha Sood, Stephen Stuart, Nicolas Fauchereau, Chris Brandolino, Ben Noll, Tristan Meyers:** Conceptualization, Methodology, Reviewing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgements

The ERA5 reanalysis data used is publicly available from Copernicus Climate Change Service (C3S):https://cds.climate.copernicus.eu/cdsapp#!/dataset/reanalysis-era5-pressure-levels.

The VCSN data is developed and maintained by NIWA and can be obtained through a data access agreement through correspondence with the authors.

The CNN models were trained and tuned using the Python deep learning API Keras which is available from https://keras.io/. The Grad-CAM interpretable deep learning code is available from: https://github.com/sicara/tf-explain.

This study was supported through the New Zealand Ministry of Business Innovation and Employment (MBIE) Strategic Science Investment Fund, and through MBIE Endeavour contract C01X2202.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.wace.2022.100525.

References

Baño-Medina, J., Manzanas, R., Gutiérrez, J.M., 2020. Configuration and intercomparison of deep learning neural models for statistical downscaling. Geosci. Model Dev. (GMD) 13 (4), 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. Clim. Dynam. https://doi.org/10.1007/s00382-021-05847-0.

Barnes, E.A., Toms, B., Hurrell, J.W., Ebert-Uphoff, I., Anderson, C., Anderson, D., 2020. Indicator patterns of forced change learned by an artificial neural network. J. Adv. Model. Earth Syst. 12 (9), e2020MS002195 https://doi.org/10.1029/2020MS002195.

Benestad, R., 2004. Empirical-statistical downscaling in climate modeling. Trans. Am Geophys. Union. 85 (42), 417–422.

Cannon, A.J., 2008. Probabilistic multisite precipitation downscaling by an expanded Bernoulli?gamma density network. J. Hydrometeorol. 9 (6), 1284–1300. https://doi.org/10.1175/2008jhm960.1.

Chapman, W.E., Subramanian, A.C., Delle Monache, L., Xie, S.P., Ralph, F.M., 2019. Improving atmospheric river forecasts with machine learning. Geophys. Res. Lett. 46 (17–18), 10627–10635. https://doi.org/10.1029/2019GL083662.

Chen, S.-T., Yu, P.-S., Tang, Y.-H., 2010. Statistical downscaling of daily precipitation using support vector machines and multivariate analysis. J. Hydrol. 385 (1–4), 13–22.

Davenport, F.V., Diffenbaugh, N.S., 2021. Using machine learning to analyze physical causes of climate change: a case study of U.S. Midwest extreme precipitation. Geophys. Res. Lett. 48 (15), e2021GL093787 https://doi.org/10.1029/2021GL093787.

Deser, C., 2020. Certain uncertainty: the role of internal climate variability in projections of regional climate change and risk management. Earth's Future 8 (12), e2020EF001854.

Ebert-Uphoff, I., Hilburn, K., 2020. Evaluation, tuning, and interpretation of neural networks for working with images in meteorological applications. Bull. Am. Meteorol. Soc. 101 (12), E2149–E2170. https://doi.org/10.1175/bams-d-20-0097.1.

Fiedler, T., Pitman, A.J., Mackenzie, K., Wood, N., Jakob, C., Perkins-Kirkpatrick, S.E., 2021. Business risk and the emergence of climate analytics. Nat. Clim. Change 11 (2), 87–94. https://doi.org/10.1038/s41558-020-00984-6.

Gibson, P.B., Chapman, W.E., Altinok, A., Delle Monache, L., DeFlorio, M.J., Waliser, D. E., 2021. Training machine learning models on climate model output yields skillful interpretable seasonal precipitation forecasts. Commun. Earth Environ. 2 (1), 159. https://doi.org/10.1038/s43247-021-00225-4.

Gibson, P.B., Waliser, D.E., Lee, H., Tian, B., Massoud, E., 2019. Climate model evaluation in the presence of observational uncertainty: precipitation indices over the contiguous United States. J. Hydrometeorol. 20 (7), 1339–1357. https://doi.org/10.1175/jhm-d-18-0230.1.

Giorgi, F., Jones, C., Asrar, G.R., 2009. Addressing climate information needs at the regional level: the CORDEX framework. World Meteorol. Organ. Bull. 58 (3), 175.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014. Generative adversarial nets. Adv. Neural Inf. Process. Syst. 27.

Gutiérrez, J.M., Maraun, D., Widmann, M., Huth, R., Hertig, E., Benestad, R., Roessler, O., Wibig, J., Wilcke, R., Kotlarski, S., San Martín, D., Herrera, S., Bedia, J., Casanueva, A., Manzanas, R., Iturbide, M., Vrac, M., Dubrovsky, M., Ribalaygua, J., Pórtoles, J., Räty, O., Räisänen, J., Hingray, B., Raynaud, D., Casado, M.J., Ramos, P., Zerenner, T., Turco, M., Bosshard, T., Štěpánek, P., Bartholy, J., Pongracz, R., Keller, D.E., Fischer, A.M., Cardoso, R.M., Soares, P.M.M., Czernecki, B., Pagé, C., 2019. An intercomparison of a large ensemble of statistical downscaling methods over Europe: results from the VALUE perfect predictor cross-validation experiment. Int. J. Climatol. 39 (9), 3750–3785. https://doi.org/10.1002/joc.5462.

Ham, Y.-G., Kim, J.-H., Luo, J.-J., 2019. Deep learning for multi-year ENSO forecasts. Nature 573 (7775), 568–572.

Hayatbini, N., Badrinath, A., Chapman, W., Delle Monache, L., Cannon, F., Gibson, P., Subramanian, A., Ralph, F.M., 2021. A two-stage deep learning framework to improve short range rainfall prediction. In: 101st American Meteorological Society Annual Meeting.

Hayatbini, N., Kong, B., Hsu, K.-l., Nguyen, P., Sorooshian, S., Stephens, G., Fowlkes, C., Nemani, R., Ganguly, S., 2019. Conditional generative adversarial networks (cGANs) for near real-time precipitation estimation from multispectral GOES-16 satellite imageries—PERSIANN-cGAN. Rem. Sens. 11 (19), 2193. https://www.mdpi.com/2072-4292/11/19/2193.

He, K., Zhang, X., Ren, S., Sun, J., 2016b. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.

He, X., Chaney, N.W., Schleiss, M., Sheffield, J., 2016a. Spatial downscaling of precipitation using adaptable random forests. Water Resour. Res. 52 (10), 8217–8237. https://doi.org/10.1002/2016WR019034.

Hersbach, H., Bell, B., Berrisford, P., Hirahara, S., Horányi, A., Muñoz-Sabater, J., Nicolas, J., Peubey, C., Radu, R., Schepers, D., 2020. The ERA5 global reanalysis. Q. J. R. Meteorol. Soc. 146 (730), 1999–2049.

Höhlein, K., Kern, M., Hewson, T., Westermann, R., 2020. A comparative study of convolutional neural network models for wind field downscaling. Meteorol. Appl. 27 (6), e1961 https://doi.org/10.1002/met.1961.

Ibbitt, R.P., Henderson, R.D., Copeland, J., Wratt, D.S., 2000. Simulating mountain runoff with meso-scale weather model rainfall estimates: a New Zealand experience. J. Hydrol. 239 (1), 19–32. https://doi.org/10.1016/S0022-1694(00)00351-6.

Jacob, D., Petersen, J., Eggert, B., Alias, A., Christensen, O.B., Bouwer, L.M., Braun, A., Colette, A., Déqué, M., Georgievski, G., Georgopoulou, E., Gobiet, A., Menut, L., Nikulin, G., Haensler, A., Hempelmann, N., Jones, C., Keuler, K., Kovats, S., Kröner, N., Kotlarski, S., Kriegsmann, A., Martin, E., van Meijgaard, E., Moseley, C., Pfeifer, S., Preuschmann, S., Radermacher, C., Radtke, K., Rechid, D., Rounsevell, M., Samuelsson, P., Somot, S., Soussana, J.-F., Teichmann, C., Valentini, R., Vautard, R., Weber, B., Yiou, P., 2014. EURO-CORDEX: new high-resolution climate change projections for European impact research. Reg. Environ. Change 14 (2), 563–578. https://doi.org/10.1007/s10113-013-0499-2.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596 (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

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

Kingma, D.P., Ba, J., 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint arXiv:1412.6980*.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., Müller, K.-R., 2019. Unmasking Clever Hans predictors and assessing what machines really learn. Nat. Commun. 10 (1), 1096. https://doi.org/10.1038/s41467-019-08987-4.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. https://doi.org/10.1038/nature14539.

Mayer, K.J., Barnes, E.A., 2021. Subseasonal forecasts of opportunity identified by an explainable neural network. Geophys. Res. Lett. 48 (10), e2020GL092092.

Maraun, D., Widmann, M., 2018. Statistical Downscaling and Bias Correction for Climate Research. Cambridge University Press.

Marotzke, J., Jakob, C., Bony, S., Dirmeyer, P.A., O'Gorman, P.A., Hawkins, E., Perkins-Kirkpatrick, S., Quéré, C.L., Nowicki, S., Paulavets, K., Seneviratne, S.I., Stevens, B., Tuma, M., 2017. Climate research must sharpen its view. Nat. Clim. Change 7 (2), 89–91. https://doi.org/10.1038/nclimate3206.

McGovern, A., Lagerquist, R., John Gagne, D., Jergensen, G.E., Elmore, K.L., Homeyer, C. R., Smith, T., 2019. Making the black box more transparent: understanding the physical implications of machine learning. Bull. Am. Meteorol. Soc. 100 (11), 2175–2199. https://doi.org/10.1175/bams-d-18-0195.1.

Mirza, M., Osindero, S., 2014. Conditional Generative Adversarial Nets. *arXiv preprint arXiv:1411.1784*.

Prabhat, Kashinath, K., Mudigonda, M., Kim, S., Kapp-Schwoerer, L., Graubner, A., Karaismailoglu, E., von Kleist, L., Kurth, T., Greiner, A., Mahesh, A., Yang, K., Lewis, C., Chen, J., Lou, A., Chandran, S., Toms, B., Chapman, W., Dagon, K., Shields, C.A., O'Brien, T., Wehner, M., Collins, W., 2021. ClimateNet: an expert-labeled open dataset and deep learning architecture for enabling high-precision analyses of extreme weather. Geosci. Model Dev. (GMD) 14 (1), 107–124. https://doi.org/10.5194/gmd-14-107-2021.

Prince, H.D., Cullen, N.J., Gibson, P.B., Conway, J., Kingston, D.G., 2021. A climatology of atmospheric rivers in New Zealand. J. Clim. 34 (11), 4383–4402. https://doi.org/10.1175/jcli-d-20-0664.1.

Rampal, N., Davies, R., 2020. On the factors that determine boundary layer albedo. J. Geophys. Res. Atmos. 125 (15), e2019JD032244 https://doi.org/10.1029/2019JD032244.

Rasp, S., Dueben, P.D., Scher, S., Weyn, J.A., Mouatadid, S., Thuerey, N., 2020. WeatherBench: a benchmark data set for data-driven weather forecasting. J. Adv. Model. Earth Syst. 12 (11), e2020MS002203 https://doi.org/10.1029/2020MS002203.

Rasp, S., Thuerey, N., 2021. Data-driven medium-range weather prediction with a Resnet pretrained on climate simulations: a new model for WeatherBench. J. Adv. Model. Earth Syst. 13 (2), e2020MS002405 https://doi.org/10.1029/2020MS002405.

Ravuri, S., Lenc, K., Willson, M., Kangin, D., Lam, R., Mirowski, P., Fitzsimons, M., Athanassiadou, M., Kashem, S., Madge, S., Prudden, R., Mandhane, A., Clark, A., Brock, A., Simonyan, K., Hadsell, R., Robinson, N., Clancy, E., Arribas, A., Mohamed, S., 2021. Skilful precipitation nowcasting using deep generative models of radar. Nature 597 (7878), 672–677. https://doi.org/10.1038/s41586-021-03854-z.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., Prabhat, 2019. Deep learning and process understanding for data-driven Earth system science. Nature 566 (7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1.

Reid, K.J., Rosier, S.M., Harrington, L.J., King, A.D., Lane, T.P., 2021. Extreme rainfall in New Zealand and its association with atmospheric rivers. Environ. Res. Lett. 16 (4), 044012 https://doi.org/10.1088/1748-9326/abeae0.

Renwick, J.A., Mullan, A.B., Porteous, A., 2009. Statistical downscaling of New Zealand climate. Weather Clim. 29, 24–44. https://doi.org/10.2307/26169704.

Ruti, P.M., Somot, S., Giorgi, F., Dubois, C., Flaounas, E., Obermann, A., Dell?Aquila, A., Pisacane, G., Harzallah, A., Lombardi, E., Ahrens, B., Akhtar, N., Alias, A., Arsouze, T., Aznar, R., Bastin, S., Bartholy, J., B?ranger, K., Beuvier, J., Bouffies-Cloch, S., Brauch, J., Cabos, W., Calmanti, S., Calvet, J.-C., Carillo, A., Conte, D., Coppola, E., Djurdjevic, V., Drobinski, P., Elizalde-Arellano, A., Gaertner, M., Galn, P., Gallardo, C., Gualdi, S., Goncalves, M., Jorba, O., Jord, G., Heveder, L., B, Lebeaupin-Brossier, C., Li, L., Liguori, G., Lionello, P., Maci?s, D., Nabat, P., nol, B., Raikovic, B., Ramage, K., Sevault, F., Sannino, G., Struglia, M.V., Sanna, A., Torma, C., Vervatis, V., 2016. Med-CORDEX initiative for mediterranean climate studies. Bull. Am. Meteorol. Soc. 97 (7), 1187–1208. https://doi.org/10.1175/bams-d-14-00176.1.

Salvi, K., Ghosh, S., Ganguly, A.R., 2016. Credibility of statistical downscaling under nonstationary climate. Clim. Dynam. 46, 1991–2023. https://doi.org/10.1007/s00382-015-2688-9.

Schoof, J.T., Pryor, S.C., 2001. Downscaling temperature and precipitation: a comparison of regression-based methods and artificial neural networks. Int. J. Climatol. 21 (7), 773–790. https://doi.org/10.1002/joc.655.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., Chen, Y., Lillicrap, T., Hui, F., Sifre, L., van den Driessche, G., Graepel, T., Hassabis, D., 2017. Mastering the game of Go without human knowledge. Nature 550 (7676), 354–359. https://doi.org/10.1038/nature24270.

Sinclair, M.R., 2002. Extratropical transition of southwest pacific tropical cyclones. Part I: climatology and mean structure changes. Mon. Weather Rev. 130 (3), 590–609. https://doi.org/10.1175/1520-0493(2002)130<0590:Etospt>2.0.Co;2.

Stengel, K., Glaws, A., Hettinger, D., King, R.N., 2020. Adversarial super-resolution of climatological wind and solar data. Proc. Natl. Acad. Sci. USA 117 (29), 16805–16815. https://doi.org/10.1073/pnas.1918964117.

Sun, L., Lan, Y., 2021. Statistical downscaling of daily temperature and precipitation over China using deep learning neural models: localization and comparison with other methods. Int. J. Climatol. 41 (2), 1128–1147.

Tait, A., Fitzharris, B., 1998. Relationships between New Zealand rainfall and south-west Pacific pressure patterns. Int. J. Climatol.: J. Royal Meteorol. Soc. 18 (4), 407–424.

Tait, A., Henderson, R., Turner, R., Zheng, X., 2006. Thin plate smoothing spline interpolation of daily rainfall for New Zealand using a climatological rainfall surface. Int. J. Climatol. 26 (14), 2097–2115. https://doi.org/10.1002/joc.1350.

Tait, A., Sturman, J., Clark, M., 2012. An assessment of the accuracy of interpolated daily rainfall for New Zealand. J. Hydrol. 25–44.

Toms, B.A., Barnes, E.A., Ebert-Uphoff, I., 2020. Physically interpretable neural networks for the geosciences: applications to Earth system variability. J. Adv. Model. Earth Syst. 12 (9), e2019MS002002 https://doi.org/10.1029/2019MS002002.

Vandal, T., Kodra, E., Ganguly, A.R., 2019. Intercomparison of machine learning methods for statistical downscaling: the case of daily and extreme precipitation. Theor. Appl. Climatol. 137 (1), 557–570. https://doi.org/10.1007/s00704-018-2613-3.

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., Ganguly, A.R., 2017. Deepsd: generating high resolution climate change projections through single image super-resolution. In: Proceedings of the 23rd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining.

Vaughan, A., Tebbutt, W., Hosking, J.S., Turner, R.E., 2021. Convolutional conditional neural processes for local climate downscaling. Geosci. Model Dev. Discuss. (GMDD) 2021, 1–25. https://doi.org/10.5194/gmd-2020-420.

Watson-Parris, D., 2021. Machine learning for weather and climate are worlds apart. Philosophical Transactions of the Royal Society A 379 (2194), 20200098.

Wilby, R.L., Wigley, T.M.L., 1997. Downscaling general circulation model output: a review of methods and limitations. Prog. Phys. Geogr.: Earth Environ. 21 (4), 530–548. https://doi.org/10.1177/030913339702100403.

Wilby, R.L., Wigley, T.M.L., Conway, D., Jones, P.D., Hewitson, B.C., Main, J., Wilks, D. S., 1998. Statistical downscaling of general circulation model output: a comparison of methods. Water Resour. Res. 34 (11), 2995–3008. https://doi.org/10.1029/98WR02577.

Xie, S.-P., Deser, C., Vecchi, G.A., Collins, M., Delworth, T.L., Hall, A., Hawkins, E., Johnson, N.C., Cassou, C., Giannini, A., 2015. Towards predictive understanding of regional climate change. Nat. Clim. Change 5 (10), 921–930.

Yang, C., Wang, N., Wang, S., Zhou, L., 2018. Performance comparison of three predictor selection methods for statistical downscaling of daily precipitation. Theor. Appl. Climatol. 131 (1), 43–54. https://doi.org/10.1007/s00704-016-1956-x.