



Downscaling long lead time daily rainfall ensemble forecasts through deep learning

Huidong Jin¹ · Weifan Jiang^{1,2} · Minzhe Chen² · Ming Li³ · K. Shuvo Bakar^{1,4} · Quanxi Shao³

Accepted: 5 April 2023 / Published online: 14 April 2023
© The Author(s) 2023

Abstract

Skilful and localised daily weather forecasts for upcoming seasons are desired by climate-sensitive sectors. Various General circulation models routinely provide such long lead time ensemble forecasts, also known as seasonal climate forecasts (SCF), but require downscaling techniques to enhance their skills from historical observations. Traditional downscaling techniques, like quantile mapping (QM), learn empirical relationships from pre-engineered predictors. Deep-learning-based downscaling techniques automatically generate and select predictors but almost all of them focus on simplified situations where low-resolution images match well with high-resolution ones, which is not the case in ensemble forecasts. To downscale ensemble rainfall forecasts, we take a two-step procedure. We first choose a suitable deep learning model, very deep super-resolution (VDSR), from several outstanding candidates, based on an ensemble forecast skill metric, continuous ranked probability score (CRPS). Secondly, via incorporating other climate variables as extra input, we develop and finalise a very deep statistical downscaling (VDSD) model based on CRPS. Both VDSR and VDSD are tested on downscaling 60 km rainfall forecasts from the Australian Community Climate and Earth-System Simulator Seasonal model version 1 (ACCESS-S1) to 12 km with lead times up to 217 days. Leave-one-year-out testing results illustrate that VDSD has normally higher forecast accuracy and skill, measured by mean absolute error and CRPS respectively, than VDSR and QM. VDSD substantially improves ACCESS-S1 raw forecasts but does not always outperform climatology, a benchmark for SCFs. Many more research efforts are required on downscaling and climate modelling for skilful SCFs.

Keywords Statistical downscaling · Ensemble forecast · Seasonal climate forecast · Deep learning · Convolutional neural network

Abbreviations

ACCESS	Australian Community Climate and Earth-System Simulator
ACCESS-S1	ACCESS Seasonal Model 1
ACCESS-R	ACCESS Regional
BI	Bicubic interpolation
BARRA	Bureau's Atmospheric High-Resolution Regional Reanalysis for Australia
CCAM	Conformal cubic atmospheric model
CNN	Convolutional neural networks
CRPS	Continuous ranked probability score
ECPP	Enhanced copula post-processing
ESRGAN	Enhanced super-resolution generative adversarial network
GCM	Global circulation model
KNN	K-nearest neighbours
QM	Quantile mapping

RCAN	Residual channel attention networks
SISR	Single image super-resolution
SCF	Seasonal climate forecast
VDSR	Very deep super-resolution
VDSD	Very deep statistical downscaling
ZG	Geopotential height

1 Introduction

Seasonal climate forecasts (SCF) can provide great value to many socioeconomic sectors such as agriculture, construction, mining, tourism, energy, and health (Merryfield et al. 2020; Manzanos 2020). As in the driest inhabited continent, e.g., Australia's agricultural system is heavily dependent on rainfall. The potential annual value added from skilful SCFs for the whole of Australia could be around A\$1.6 billion for the agricultural sector and A\$192

Extended author information available on the last page of the article

million for the construction sector (The Centre for International Economics 2014). Thus, various ensemble SCFs based on General Circulation Models (GCMs) are routinely produced around the world (Hudson et al. 2017; Merryfield et al. 2020; Saha et al. 2014; Johnson et al. 2019). Despite ongoing development in GCMs, their typical grid resolutions (~ 100 km) still limit their direct application to weather-sensitive sectors (Baño-Medina et al. 2020; Schepen et al. 2020; Kusunose and Mahmood 2016; Luo 2016). The barriers could be overcome via downscaling techniques which generate more skilful and localised forecasts by making use of local observations (Maraun and Widmann 2018; Bettolli et al. 2021).

Due to the challenges caused by spatial-temporal variabilities of climate variables, especially precipitation, a large number of downscaling techniques have been developed, including dynamical downscaling (Ratnam et al. 2017; Thatcher and McGregor 2009; Luo 2016), statistical downscaling (Maraun and Widmann 2018), and the recent development of deep-learning-based downscaling. Dynamical downscaling uses a physics-based climate model, such as the conformal cubic atmospheric model (CCAM) (Thatcher and McGregor 2009), forced by boundary conditions from a GCM, to simulate atmospheric conditions at a finer resolution. Statistical downscaling builds empirical relationships between GCM raw hindcasts and historical observations and then uses them to remove systematic biases, adjust the uncertainty spread, and restore local daily variability of raw forecasts (Ahmadalipour et al. 2018; Maraun and Widmann 2018; Şan et al. 2022; Shao and Li 2013; Crimp et al. 2019). A typical example is quantile mapping (QM) which assumes that the distribution of model-simulated data at a given location should preserve the distribution of observed data (Michelangeli et al. 2009; Li and Jin 2020). Comparisons between traditional statistical and dynamical downscaling suggest that neither group of methods are superior, however, in practice, computationally cheaper statistical methods are widely used (Baño-Medina et al. 2020). Pre-engineered predictors and relationships limit these statistical downscaling techniques to exploit various spatio-temporal dependencies, and then their abilities to capture information beyond prior knowledge (Baño-Medina et al. 2020; Liu et al. 2020). Automatic feature extraction and selection integrated into the modelling process with deep learning, especially convolutional neural networks (CNNs), have achieved notable success in modelling data with spatial context, recently in climate science (Reichstein et al. 2019). Deep learning has been successfully used in precipitation nowcasting (Shi et al. 2015; Espenholt et al. 2022), which predicts rainfall intensity in a region over 3–6 h, and precipitation parameterisations from GCMs (Pan et al. 2019). More related to this study, several downscaling

techniques have been developed based on single image super-resolution (SISR) techniques. For example, DeepSD was proposed by augmenting multi-scale topography into stacked super-resolution convolutional neural networks (SRCNN) (Vandal et al. 2017). DeepSD succeeded in downscaling daily rainfall data to 12.5 km. For long-term climate projection, Rodrigues et al. (2018) proposed a very deep CNN-based SISR strategy to interpolate low-resolution 125 km weather data to 25 km output for weather forecasts. Baño-Medina et al. (2020) assessed CNN methods with three convolutional layers followed by different connection layers for downscaling 200 km reanalysis precipitation to 50 km observational grids over Europe. Super-resolution deep residual network (SRDRN) was proposed based on a deep CNN with residual blocks and batch normalisation for downscaling daily precipitation and temperature (Wang et al. 2021). It leaves behind bias-correction. Liu et al. (2020) presented YNet which consists of an encoder–decoder-like architecture with residual learning through skip connections and fusion layers to enable the incorporation of topological and climatological data as auxiliary inputs for downscaling. It was tested on monthly precipitation means that have different characteristics from daily data.

Downscaling SCFs looks similar to SISR as both aim at getting high-resolution images from low-resolution ones if climate variable data are treated as images (Liu et al. 2020). However, there are several differences.

1. Inputs and outputs in downscaling SCFs are from different sources, such as low-resolution forecasts from GCM vs historical weather data (Liu et al. 2020). In SISR, the low-resolution images and high-resolution target ones are arguably from the same source, e.g., the high-resolution images are often aggregated to form low-resolution images as the training inputs (Wang et al. 2020). As far as we know, almost all the deep-learning-based downscaling techniques focused on such a simplified situation (Vandal et al. 2017; Rodrigues et al. 2018; Liu et al. 2020; Wang et al. 2021).
2. Bias and displacement in space or time are common in SCFs, especially for precipitation, due to the inherent complexity of our climate system. To mitigate these mismatch issues, multiple possible forecast trajectories are provided as a practical standard for short or long lead time forecasts (Hudson et al. 2017; Merryfield et al. 2020; Johnson et al. 2019). Therefore, downscaling performance should be evaluated in terms of both forecast accuracy and overall ensemble forecast skill by considering forecast uncertainty (Grimm et al. 2006; Li and Jin 2020; Kusunose and Mahmood 2016). The latter is predominant in the literature (Grimm et al.

2006; Ferro et al. 2008; Schepen et al. 2020) but, as far as we know, has never been used in deep-learning-based downscaling.

- Downscaling precipitation can use auxiliary variables (Maraun and Widmann 2018; Bettolli et al. 2021). Rainfall events are often associated with other climate variables, e.g., intense low-pressure systems (Pan et al. 2019; Baño-Medina et al. 2020; Liu et al. 2020), which are found often beneficial for downscaling (Baño-Medina et al. 2020; Liu et al. 2020).

To address these differences, we model downscaling ensemble forecasts as a SISR problem with an additional target on maximising ensemble forecast skills. To leverage advanced deep learning techniques, for downscaling long lead time daily precipitation forecasts for the whole of Australia (Sect. 2), we choose very deep super-resolution (VDSR) (Kim et al. 2016) from outstanding SISR techniques as a suitable candidate for downscaling based on the continuous ranked probability score (CRPS), a widely used ensemble forecast skill metric (Grimit et al. 2006; Li and Jin 2020; Ferro et al. 2008; Schepen et al. 2020). Raw precipitation forecasts from GCMs are partially parameterised and are usually considered less reliable compared to directly resolved variables, such as pressure and temperature (Pan et al. 2019). To improve its downscaling performance, we incorporate other resolved climate variables into VDSR and propose a very deep statistical downscaling (VDSD) model. The VDSD structure is again finalised based on CRPS (Sect. 3). It is tested on real-world application scenarios for downscaling 60 km SCFs to 12 km. Leave-one-year-out cross-validation results illustrate its better performance than VDSR and two classical downscaling techniques in terms of both forecast accuracy and ensemble forecast skills. In addition, its performance is better than or comparable with climatology, a benchmark for long lead time climate forecasts. VDSD does not always outperform climatology (Sect. 4). Many more research efforts are required on downscaling and climate modelling for skilful SCFs.

2 Data and pre-processing

2.1 ACCESS-S1 forecast and calibrated data

This work focuses on downscaling daily rainfall forecasts for the whole of Australia. We use daily rainfall retrospective forecasts from Australia's operational seasonal climate forecast system, the Australian Community Climate and Earth-System Simulator Seasonal model version 1 (ACCESS-S1) (Hudson et al. 2017; Bureau National Operations Centre 2019), which is used for climate

outlooks on multi-week through to seasonal timescales. Its atmospheric model has enhancements to the ensemble generation strategy to make it appropriate for sub-seasonal forecasting and large ensembles. The resolution of the atmospheric model is raised to 0.6° , nearly 60 km in the mid-latitudes. The hindcast data¹ of ACCESS-S1, from 1990 to 2012, are publicly available.² Within each year, it has forecasts on 48 different initialisation dates (i.e. 1st, 9th, 17th, and 25th of each calendar month). Its forecasts have 11 ensemble members, each of which provides a full description of the evolution of weather for the upcoming 217 days. Daily precipitation data from ACCESS-S1 are based on the BoM's day definition of 9 am to 9 am (local time). Three precipitation forecasts for 7 Jan 2012 made on 1 Jan 2012 with a lead time of 6 days are illustrated in the second column of Fig. 1.

ACCESS-S1 data also provides a calibrated version. For each forecast initialisation date, lead time, and grid point location, it has a calibrated function to downscale to a 5 km resolution (Bureau National Operations Centre 2019). For a given forecast day, the calibration functions first carry out spatial interpolation using bilinear interpolation to high spatial resolution and then apply QM to adjust the bias and spread between observations and forecasts in the other 22 years. Bilinear interpolation is to interpolate an image using repeated linear interpolation. It first linearly interpolates a low-resolution image in one direction, and then in the second direction. QM downscaling for a location can be formulated as $x^{(QM)} = F_o^{-1}(F_f(x_f))$ where F_o^{-1} is the inverse function of F_o , and F_f and F_o indicate the cumulative distribution functions (CDFs, aka quantile functions) of raw forecasts x_f and observations x_o respectively (Maraun and Widmann 2018). The empirical distributions of raw forecasts and observations over a 15-days reference period are used as the estimates of F_f and F_o (Li and Jin 2020; Bureau National Operations Centre 2019). We use the calibrated data for forecast skill comparison, denoted as QM hereafter.

2.2 BARRA reanalysis data

Bureau's Atmospheric High-Resolution Regional Reanalysis for Australia (BARRA),³ is a regional numerical climate forecast model using the Australian Community Climate and Earth-System Simulator-Regional (ACCESS-R), Australia's first reanalysis model of the atmosphere (Su et al. 2019). Through assimilating local surface observations and locally derived wind vectors that are not available

¹ We call these 'forecast' hereafter in the paper for simplicity.

² <http://www.bom.gov.au/research/projects/ACCESS-S/>.

³ BARRA data are available from <http://www.bom.gov.au/research/projects/reanalysis/>.

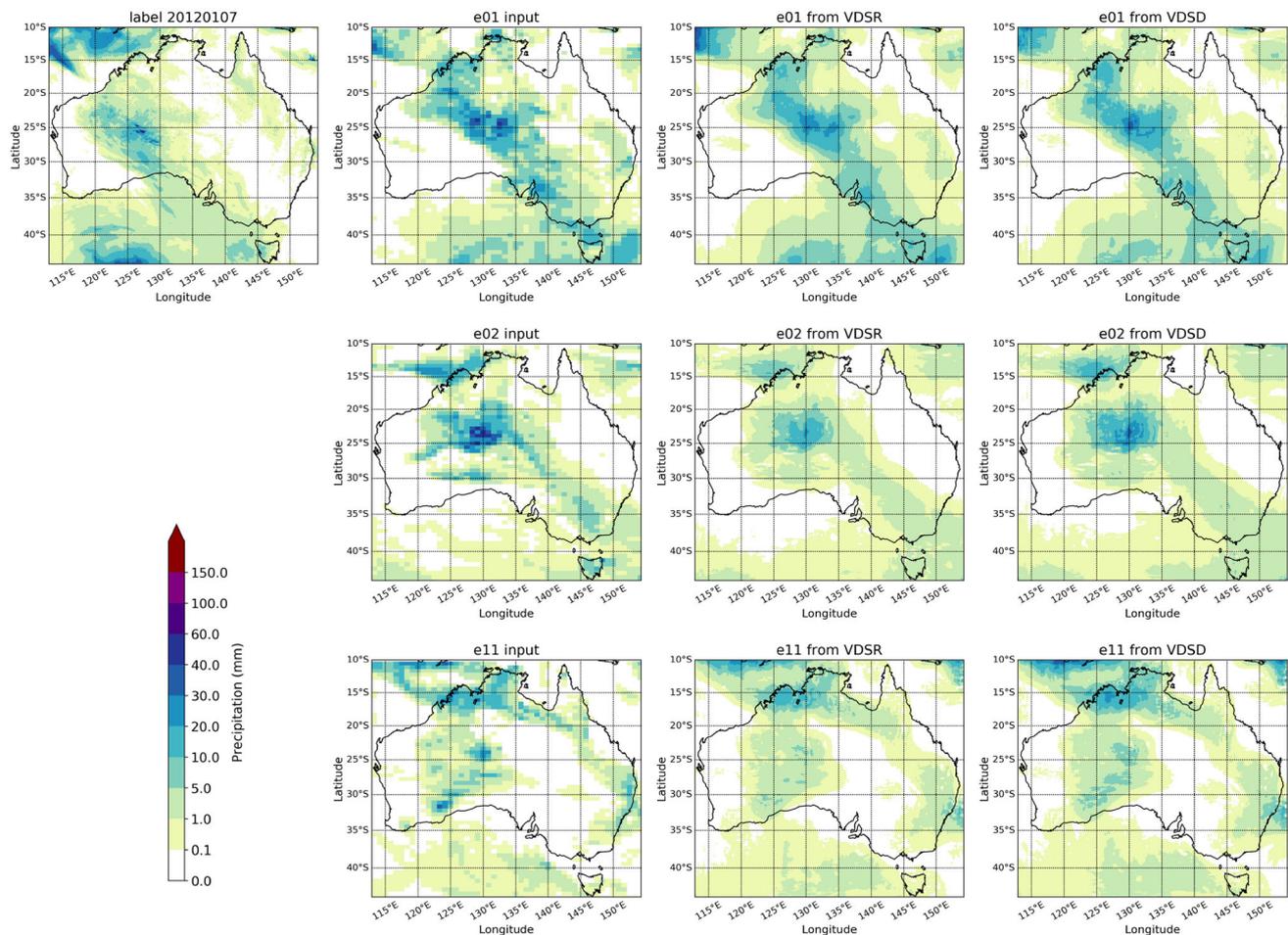


Fig. 1 Reanalysis data and daily rainfall ensemble forecasts for 7 Jan 2012 with a lead time of 6 days for the forecasts made on 1 Jan 2012. Images in the four columns are the high-resolution image from BARRA reanalysis data, ensemble member forecasts from ACCESS-

to global reanalysis models, BARRA reaches a good trade-off between the spatial resolution and consistency with precipitation observations (Acharya et al. 2019). Its spatial resolution of 0.12° , is realised in the whole region of Australia and New Zealand. Six-hour accumulated precipitation, obtained from BARRA from 1 Jan 1990 to 31 Dec 2013, is aggregated to daily frequency by taking the sum of the four 6-h grid point values within each 9 am to 9 am 24-h window.

2.3 Pre-processing

We choose a region from 9° S to 43.7425° S and 112.9° E to 154.25° E as our study region, covering all the Australian landmass (see, Fig. 1). As pre-processing, we crop all the climate variable surfaces to the same area defined in the case study region. These climate variables have

S1 after bicubic interpolation, and downscaled results from VDSR and VDSD respectively. Only the 1st, 2nd, and 11th members are illustrated in the figure

different value ranges. For example, precipitation ranges from 0 to 900 mm per day, and geopotential height at 850 hPa ranges from 1200 to 1600 m. To bring climate variables to have the same value range of $[0, 1]$ during learning, we carry out simple linear normalisation.

To facilitate 4-time image super-resolution, we generate two versions of ACCESS-S1 forecasts via bicubic interpolation (BI). One is 12 km, used as inputs for VDSR and our proposed VDSD, and the second is 48 km, used as inputs for other deep learning models. We pair the ACCESS-S1 forecasts made on date i with forecast lead time l day with the BARRA reanalysis data on date $d(=i+l)$ together for training/validation/test. There are about 2.62 million image pairs for each spatial resolution. To save training time, we only use the first seven lead time forecast pairs for each initialisation date in the training period.

3 Methodology

By treating ACCESS-S1 raw forecast and BARRA reanalysis daily data as low- and high-resolution images respectively, we model our SCF downscaling problem as a single image super-resolution (SISR) problem with an additional target on maximising ensemble forecast skill to leverage advanced deep learning techniques and mitigate the mismatch issues. We first briefly describe and review SISR techniques.

3.1 Image super-resolution and deep learning

SISR is to recover a high-resolution image H_a from a low-resolution one L_a . L_a is often regarded as the result of a degradation function $L_a = \mathcal{D}(H_a; \gamma)$ with parameter γ (Wang et al. 2020). Most super-resolution data sets are obtained by aggregation or degradation mapping from high-resolution images (Wang et al. 2020). Given low- and high-resolution image pairs $\{L_a, H_a\}_{a=1, \dots, p}$, SISR is to find a super-resolution mapping function \mathcal{F} with parameters θ to generate, from low-resolution ones L_a , high resolution images S_a as close as possible to H_a

$$S_a = \mathcal{F}(L_a; \theta). \quad (1)$$

The simplest SISR techniques are spatial interpolation, such as nearest neighbour interpolation, bilinear interpolation, and BI. BI uses cubic splines or other polynomial techniques to interpolate data on a two-dimensional regular grid, which could sharpen or enlarge images. BI can consider more neighbouring grid points, and get smoother images with fewer interpolation artefacts than bilinear interpolation. BI is often considered to be the baseline for spatial downscaling of precipitation fields (Vandal et al. 2017).

Since Dong et al. (2014) first introduced super-resolution CNN (SRCNN), deep-learning-based SISR techniques have been widely developed and achieved marvellous improvements in terms of image or perceptual quality (Wang et al. 2020). Most of them are based on CNNs (Liu et al. 2020). As one of the most popular deep learning neural networks, a CNN is mainly used in processing grid-like data such as images. Besides an input and an output layer, a CNN has one or more hidden layers. The hidden layers could be of several different types: convolution, activation, pooling, normalisation, and fully-connected one. Its core building block, a convolution layer, uses multiple filters to slide across the height and width of a matrix or matrices, which is input for this layer, to generate a spatial representation of the receptive region of these filters. The spatial representation forms feature maps, which are output to feed the next layer. Multiple

convolution layers allow us to study features in quite different local areas, and re-using filters reduces the number of model parameters for training (Dong et al. 2014; Wang et al. 2020). These SISR models use several additional network design techniques, such as gradient clipping and residual learning in very deep super-resolution (Kim et al. 2016), residual dense block in Dense feature fusion (DFF) (Zhang et al. 2018b), and attention mechanism in residual channel attention network (RCAN) (Zhang et al. 2018a). RCAN achieve state-of-the-art in terms of image quality measured by peak signal-to-noise ratio (PSNR) (Wang et al. 2020). To generate more realistic images, encoder–decoder networks and generative adversarial networks (GAN), are later employed in super-resolution GAN (SRGAN) (Ledig et al. 2017) and enhanced SRGAN (ESRGAN) (Wang et al. 2018), and multiple semantic information used in (Rad et al. 2019). ESRGAN outperforms various models in terms of image perceptual quality (Wang et al. 2018).

By leveraging these SISR models for maximising ensemble forecast skills, we develop our SCF downscaling techniques in two steps. First, we select representative deep learning techniques based on their outstanding SISR performance, and train them to generate a high-resolution precipitation image from each low-resolution forecast ensemble member and then choose the one with the highest ensemble forecast skill for the whole of Australia on a separate validation data set (Sect. 3.2). Secondly, based on the selected deep learning structure, we incorporate other resolved climate variables and propose VDSD to enhance downscaling (Sect. 3.3).

3.2 Model selection for downscaling ensemble forecasts

SCFs, covering long lead times from weeks to multiple months, are located at the transition between weather forecasting and climate projection, and have been a big challenge in the weather and climate communities for years (Merryfield et al. 2020). To remedy the simplification nature of GCMs (Maraun and Widmann 2018), ensemble SCFs become an operational standard where multiple trajectories are provided on a forecast initialisation date i . For these forecasts, let $X^{(i,l,e)} \equiv \left\{ x_{j,k}^{(i,l,e)} \right\}_{m_0 \times n_0} \in \mathcal{R}^{m_0 \times n_0}$ and $\hat{Y}^{(i,l,e)} \equiv \left\{ y_{j,k}^{(i,l,e)} \right\}_{m \times n}$ be precipitation raw forecast and its associated downscaled forecast, respectively, with lead time l day, ensemble number e ($e = 1, 2, \dots, E$) for grid point (j, k) . Their associated precipitation observation for target-date $d (= i + l)$, is $\left\{ y_{j,k}^{(d)} \right\}_{m \times n}$. Thus, there are E different forecasts made on date i for date d for each

location (j, k) . For our downscaling application, $E = 11$, $m_0 = 79$, $n_0 = 94$, $m = 316$, $n = 376$, $l = 0, \dots, 216$ days. All the ensemble members target at the same date, say Jan 7 Dec 2012 in Fig. 1, and have the same target images. The forecast accuracy metrics for $\hat{Y}^{(i,l,e)}$ such as mean absolute error (MAE) $\frac{\sum_{j,k} \left\| \hat{y}_{j,k}^{(i,l,e)} - y_{j,k}^{(d)} \right\|}{\sum_{j,k} 1}$,⁴ Root mean square error

(RMSE) $\sqrt{\frac{\sum_{j,k} \left(\hat{y}_{j,k}^{(i,l,e)} - y_{j,k}^{(d)} \right)^2}{\sum_{j,k} 1}}$; and PSNR are not enough,

especially for considering possible bias and displacement within each ensemble forecast member. The continuous ranked probability score (CRPS), which generalises the MAE, is one of the most widely used overall forecast skill metrics where probabilistic or ensemble forecasts are involved. It is a surrogate measure of forecast reliability, sharpness and efficiency (Hersbach 2000). It is defined as

$$CRPS \left(\left\{ \hat{y}_{j,k}^{(i,l,e)} \right\}_{e=1, \dots, E}, y_{j,k}^{(d)} \right) = \int_{s=0}^1 \left(\hat{F}_{j,k}^{(i,l)}(s) - \mathbb{1}(s \leq y_{j,k}^{(d)}) \right)^2 ds, \quad (2)$$

where $\hat{F}_{j,k}^{(i,l)}(s)$ is an (often empirical) cumulative distribution function derived from an ensemble forecast $\left\{ \hat{y}_{j,k}^{(i,l,e)} \right\}_{e=1, \dots, E}$ and $\mathbb{1}$ is an indicator function, which represents the exceedance of the forecast compared to the actual observation $y_{j,k}^{(d)}$. It is denoted as $CRPS \left(\hat{F}_{j,k}^{(i,l)}, y_{j,k}^{(d)} \right)$ hereafter for simplicity. CRPS considers both forecast bias and forecast uncertainty of ensemble members. It reaches its minimum 0 when all the forecasts are identical with the observation, and increases with forecast bias and spread of the ensemble forecast.

As the initialisation conditions in SCFs vary from one initialisation date to another, these ensemble members do not correlate across initialisation dates. Instead of generating an aggregated forecast from an ensemble of forecasts e.g., in Liu et al. (2020), we need to generate one high-resolution forecast precipitation image from each low-resolution forecast image, such that these high-resolution forecasts can be used directly by applications, such as feeding into biophysical models (Schepen et al. 2020; Basso and Liu 2019; Luo 2016; Jin et al. 2022). Thus, our downscaling problem can be defined as follows. For low-resolution output images from GCMs, precipitation image $X^{(i,l,e)} \in \mathcal{R}^{m_0 \times n_0}$ and other climate variable images $Z^{(i,l,e)} \in \mathcal{R}^{m_0 \times n_0 \times p}$ concerning a target high-resolution image $Y^{(d)} \in \mathcal{R}^{m \times n}$, we would like to find such a function \mathcal{G} , which

generates high-resolution precipitation image as the same resolution as $Y^{(d)}$,

$$\hat{Y}^{(i,l,e)} = \left\{ \hat{y}_{j,k}^{(i,l,e)} \right\}_{m \times n} = \mathcal{G} \left(X^{(i,l,e)}, Z^{(i,l,e)}; \theta \right), \quad (3)$$

that can minimise the average CRPS across all the validation image pairs:

$$\overline{CRPS} = \frac{\sum_{i,l,j,k} n w_{j,k}^{(i,l)} CRPS \left(\left\{ \hat{y}_{j,k}^{(i,l,e)} \right\}_{e=1, \dots, E}, y_{j,k}^{(d)} \right)}{\sum_{i,l,j,k} w_{j,k}^{(i,l)}} \quad (4)$$

where $w_{j,k}^{(i,l)}$ is the weight for the ensemble forecast made on date i , lead time l at location (j, k) , and $d = i + l$. We use $w_{j,k}^{(i,l)} \equiv 1$ for this study for simplicity. The downscaling problem is modelled as a SISR problem (Eq. 3) but optimising the average ensemble forecast skill CRPS (Eq. 4). Our deep learning downscaling solution is to provide a SISR model, i.e., the function \mathcal{G} , to generate a high-resolution precipitation image $\hat{Y}^{(i,l,e)}$ from each low-resolution precipitation forecast image $X^{(i,l,e)}$, as well as other low-resolution climate variable $Z^{(i,l,e)}$ from a climate model. Thus, E high-resolution forecasts $\left\{ \hat{Y}^{(i,l,e)} \right\}_{e=1, \dots, E}$, corresponding to an ensemble of E low-resolution forecasts $\left\{ X^{(i,l,e)} \right\}_{e=1, \dots, E}$, can form a high-resolution ensemble forecast for a forecast target day $d (= i + l)$. Averaging over the CRPS of these ensemble forecasts on the validation days is used as an optimisation objective to finalise the deep learning network architecture.

To determine such a good function \mathcal{G} and its parameter θ , as illustrated in Fig. 2, we take a relatively simple two-step procedure: the first step is to find a suitable deep learning model as \mathcal{F} in Eq. 1 according to the average CRPS, and then incorporate extra variables $Z^{(i,l,e)}$ to enhance its downscaling performance. Before that, as illustrated in Fig. 2, we split the forecast data into two groups according to their initialisation dates. In the model selection and development steps in this and next subsections, we partition all the initialisation dates in the 23 years of ACCESS-S1 randomly into two groups. The first group has 1056 initialisation dates and image pairs from this group are used for the model training. The image pairs from the remaining 48 initialisation dates are left for forecast skill validation. The forecast skill CRPS is averaged over the forecasts with lead times up to 30 days. For the performance test in the next section, we put initialisation dates in 1 year out for testing and the other year data for model parameter training to facilitate a fair comparison with Climatology and QM. In the first stage of SISR model selection, we treat our downscaling problem as image super-resolution and employ three SISR models, VDSR (Kim et al. 2016), RCAN (Zhang et al. 2018a), and

⁴ As the LaTeX class sn-jnl.cls recommended by the journal does not support $\| \cdot \|$ is used to indicate absolute value like $|\cdot|$ in this manuscript.

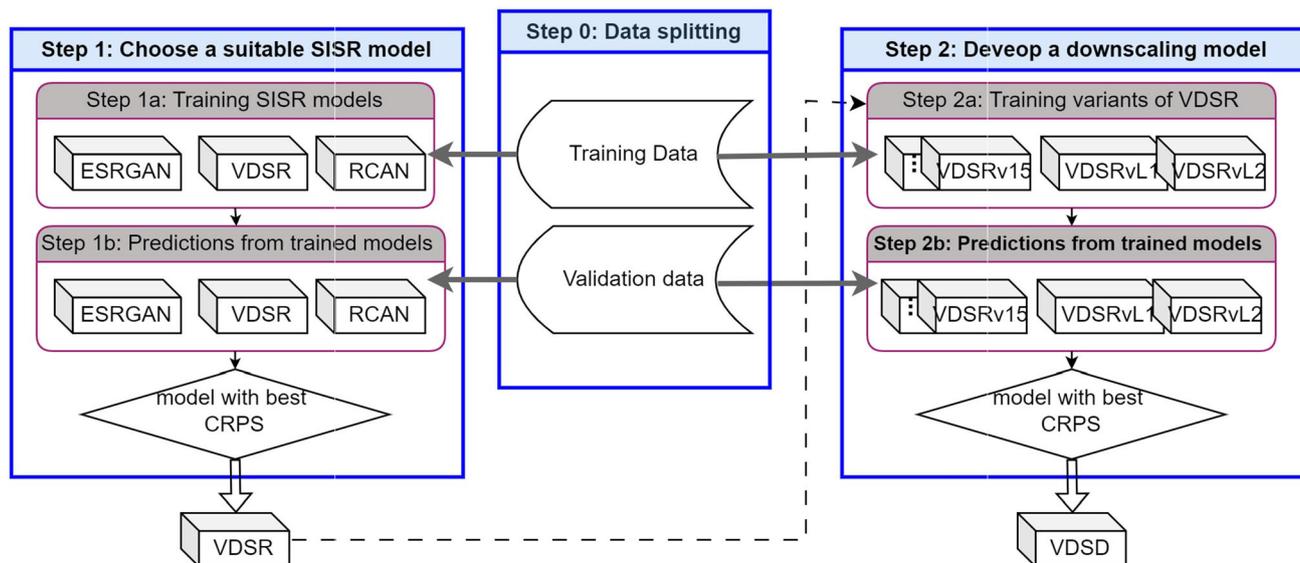


Fig. 2 Flow Chart of model selection and development based on the forecast skill CRPS on the validation data. ESRGAN, VDSR, and RCAN are three SISR deep learning models. VDSRv indicates

different variants from VDSR, such as including additional climate variables (VDSRv1 to VDSRv15), and different learning objective function like VDSRvL1 and VDSRvL2

ESRGAN (Wang et al. 2018). They are chosen because of their outstanding performance on SISR (Wang et al. 2020, 2018). The parameter learning is based on image super-resolution, and our deep learning-based problem becomes:

$$\hat{\theta} = \arg \min_{\theta} \left[\mathcal{L}(\hat{Y}^{(i,l,e)}, Y^{(d)}) + \lambda \mathcal{L}_h \right], \tag{5}$$

where \mathcal{L} is the pixel-level loss function, such as L1,

$$\mathcal{L}(\hat{Y}^{(i,l,e)}, Y^{(d)}) = \frac{\sum_{k,j} \left\| \hat{y}_{k,j}^{(i,l,e)} - y_{k,j}^{(d)} \right\|}{\sum_{k,j} 1}$$

; L2,

$$\mathcal{L}(\hat{Y}^{(i,l,e)}, Y^{(d)}) = \sqrt{\frac{\sum_{k,j} \left(\hat{y}_{k,j}^{(i,l,e)} - y_{k,j}^{(d)} \right)^2}{\sum_{k,j} 1}}$$

; or PSNR between high-resolution images $Y^{(d)}$ and super-resolution images $\hat{Y}^{(i,l,e)}$ from Eq. 3; λ is a trade-off parameter and \mathcal{L}_h is the high-level image loss like perceptual and/or adversarial loss (Wang et al. 2018).

During our model training, we used the default settings of the three SISR models. For example, VDSR used the whole images, and RCAN and ESRGAN used cropped high-resolution patches with the spatial size of 192×192 and 128×128 respectively. Their mini-batch size was 64, 16 and 16 respectively (Wang et al. 2018; Kim et al. 2016; Zhang et al. 2018a). Restricted by the computational resources we could access, the mini-batch of RCAN and ESRGAN was hard to increase. On the separate validation

data set for lead times up to 30 days, the average forecast CRPS of trained VDSR, RCAN, and ESRGAN across Australia is 1.38, 1.53, and 1.68 respectively. The forecast skill of RCAN and ESRGAN is not as good as QM with a forecast CRPS of 1.39. For our downscaling problem, both ESRGAN and RCAN were found slow to converge, partially because cross-channel dependency mechanisms became useless for our data. VDSR converged fast and outperformed QM. We selected the VDSR model for further development. We also tried a few different settings of VDSR, such as with 8, 12, 15, 20, 30, or 36 convolutional layers. Its forecast CRPS on the validation data set decreased from 8 to 20 layers, and after that, did not change much on the validation data set. We stuck with 20 layers as Kim et al. (2016) did for SISR.

3.3 Very deep statistical downscaling (VDS D)

As we discussed earlier, other climate variables such as temperature or air pressure could influence precipitation and have often been used for precipitation simulation and downscaling (Pan et al. 2019; Baño-Medina et al. 2020). To further improve downscaling performance of VDSR, we include these climate variables in our very deep statistical downscaling (VDS D). The climate variables, different from precipitation, are resolvable in climate modelling and often have more reliable forecasts (Pan et al. 2019; Baño-Medina et al. 2020; Merryfield et al. 2020). Including these climate variables and their combination into VDSR, we have 15 variants of VDSR, indicated as VDSRv1 to VDSRv15, as in Fig. 2. The overall

structure of finalised VDSR is illustrated in Fig. 3 in which Geopotential Height (ZG) at 850 hPa is incorporated as the additional input for the 9th ensemble members as in Fig. 1.

VDSR, modified from VDSR, mainly has three parts: input, intermediate feature extraction, and output layers. It can take precipitation images and other climate images, such as ZG as input in Eq. 3. These input images have been pre-processed with the same spatial resolution as high-resolution output images, and the same normalised value range (0 to 1, detailed in Sect. 2.3). These input images go through multiple feature extraction layers/blocks. These feature extraction layers have both convolution and activation modules, while the output layer only has a convolution module to generate a residual precipitation image. Adding back the interpolated raw precipitation image, indicated by the rightwards arrow with tip downwards in Fig. 3, it finally generates an output image at the same resolution as the target image. VDSR, similar to VDSR, does not have pooling and normalisation layers, as it maintains residual learning which has been widely demonstrated to contribute to robust and speedy training for SISR (Kim et al. 2016; Wang et al. 2020).

As shown in Fig. 3, two or more input images X_{lr} and Z_{lr} , which represent the raw climate forecasts after upsampling (i.e., with 48 km spatial resolution for our applications), first go through the input layer. This layer has a convolution layer and a rectified linear unit (ReLU) activation layer that forces a negative input to zero and leaves a positive input unchanged. The convolution layer has $64 \times 3 \times 3$ matrix filters that are slid across the input image and multiplied with the input image to produce 64 first-level feature images. Then the ReLU layer performs the ReLU function ($\max(0, x)$) to force negative values

from the feature images to be zero. The operation can be formulated as

$$M_0 = B(X_{lr}, Z_{lr}) = ReLU(Conv(X_{lr}, Z_{lr})) \tag{6}$$

where M_0 is the first level feature images generated by the input layer, and $ReLU()$ and $Conv()$ are ReLU and convolution layers that perform the ReLU function and 2-dimensional convolution respectively. Each convolutional layer has 64 filters and produces 64 feature images. The filter size is set to be 3×3 . Both the padding and stride step lengths are 1. Therefore, the size of each feature image is the same as the size of high-resolution images. Suppose the size of input images is $m \times n \times 2$, then the size of the feature image generated by the input layer is $m \times n \times 64$. These basic features then go through multiple intermediate blocks. The intermediate blocks are identical and each of them consists of a convolutional layer, which extracts deeper spatial features, and a ReLU layer, which introduces nonlinearity and interaction. Each convolutional layer takes 64 feature images from the previous block as input. Therefore, the operation of each intermediate block is the same, which can be written as

$$M_t = B(M_{t-1}) = ReLU(Conv(M_{t-1})) = B^t(M_0) \tag{7}$$

where M_t represents the t th level feature image, and B is the operation of an intermediate block.

The output layer is a convolutional layer that converts 64 high-level feature images into a residual image—that is to use spatial patterns discovered to predict the difference between upsampled low-resolution rainfall forecast and the target image. Finally, the residual image is added to the upsampled precipitation input image to generate a super-resolution precipitation forecast. The unknown parameters,

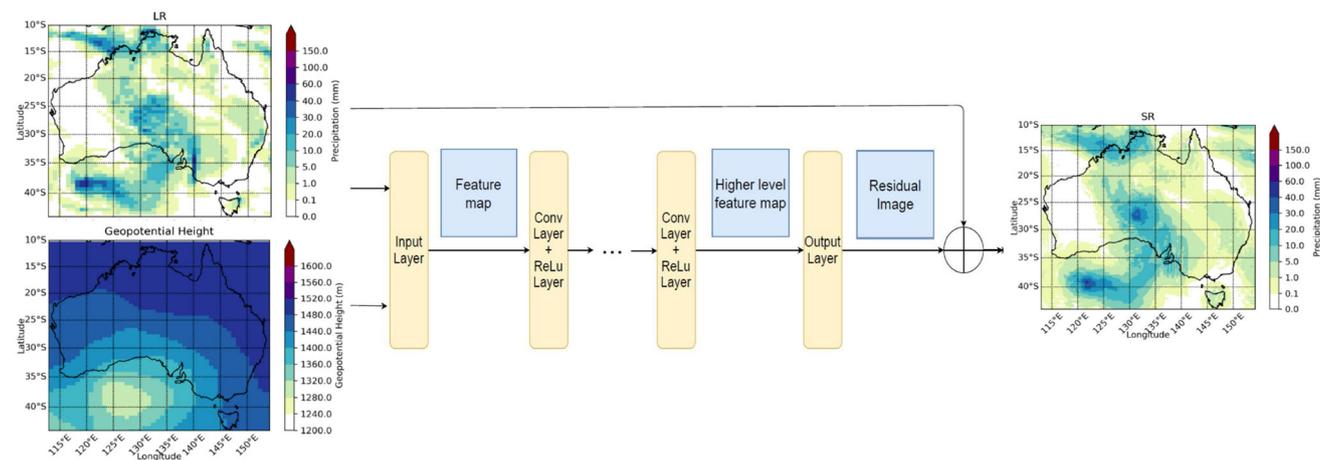


Fig. 3 The structure of the VDSR model, modified from VDSR (Kim et al. 2016), where \oplus represents element-wise matrix addition with input precipitation image, orange blocks indicate layers of the neural network, blue rectangles indicate feature images, input and output

daily rainfall data images are on the left-hand and right-hand sides respectively. For easy understanding, these input/output images are shown in the original scale, instead of the normalised scale used in downscaling

such as these in filters, will be learned from training image pairs.

A convolution layer is used in each of input, intermediate feature extraction, and output layers. A convolution layer, together with a ReLU layer, in each feature extraction block, uses 64 filters with a size of 3×3 to slide through its input images to generate different spatial features with different receptive fields. The receptive field on the original input images gradually becomes larger after passing through one feature extraction block. Around 20 feature extraction blocks can generate spatial features from quite different local areas from both precipitation and the other climate variable images. These features are tuned during model training by adjusting the parameters in the filters. These features may capture potentially complex spatial patterns, which are expected to improve downscaling.

We again used the forecast CRPS across the whole of Australia on the separate validation data set to finalise VDSM. We tested two different types of variants of VDSM. (1) One added extra input images for downscaling. We tried four climate variables from ACCESS-S1 and their combinations: ZG, daily maximum temperature, daily minimum temperature, and sea level pressure. There are total 15 variants of VDSM. Our results showed that adding more climate variables than ZG rarely improved CRPS, and often deteriorated its ensemble forecast skills. (2) The other was to try two different loss functions in Eq 5, i.e., L1 or L2. It leads to two more variants as in Fig. 2. L1 gave a slightly lower average forecast CRPS of 1.36 for lead times up to 30 days and would be used in our finalised VDSM model. Thus, there are two main differences between VDSM and VDSR. VDSM uses extra input images, and L1, instead of L2 as the learning objective.

4 Test results and discussions

To illustrate the downscaling performance of VDSR and VDSM whose network structure has been finalised in Sect. 3, we used the last 3-year retrospective forecast data for cross-validation. To facilitate a fair comparison with the ACCESS-S1 calibrated data set (i.e., the QM method), we conducted two leave-one-year-out tests. (1) We took forecasts made on the 48 initialisation dates in 2012 for testing and the other forecasts made before 2012 for training the downscaling models' parameters. Daily BARRA precipitation data between 1 Jan 2012 and 29 July 2013 were included in the test set as the ACCESS-S1 forecasts made on 25 Dec 2012 cover up to 29 July 2013 for its 217-day lead time forecasts. (2) We left forecasts made on the 48 initialisation dates in 2010 as the second test set and took ACCESS-S1 forecasts made in other

years, i.e., 1990–2009 and 2011–2012, as training data. Daily precipitation data between 1 Jan 2010 and 29 July 2011 were used in the test. In total, around 1152 daily precipitation images from BARRA were used in the prediction performance test.

4.1 Downscaling performance metrics

A benchmark for SCFs for a given year is to use observations on the same day of other years except for the target year in a base period to form an ensemble forecast, which is often called climatology (Li and Jin 2020; Schepen et al. 2020). In this study, we use 1990–2012 as the base period, and thus there are 22 ensemble members in our climatology ensemble forecasts.

As in Sect. 3.2, the average CRPS of ensemble forecasts for each grid point (k, j) on test data is treated as an overall ensemble forecast skill assessment. For each grid point, averaging across all the initialisation dates in the test year, we obtain the average CRPS of a forecast model m for a lead time l , $\overline{CRPS}_{l,k,j}^{(m)}$. For further comparison with climatology and easy understanding, we calculate the CRPS skill score for model m against the CRPS of climatology as follows.

$$CRPS_SS_{l,k,j}^{(m)} = 1 - \frac{\overline{CRPS}_{l,k,j}^{(m)}}{\overline{CRPS}_{l,k,j}^{(clim)}} \tag{8}$$

The model with a higher CRPS skill score is preferred. The skill score ranges from $-\infty$ to 1 and reaches its maximum of 1 when the \overline{CRPS} is 0, i.e., a perfect forecast where each forecast is identical to its associated observation. The skill score is zero if a forecast has the same average CRPS as climatology. A positive CRPS skill score indicates the downscaled forecast is better than the climatology model, and vice versa.

As downscaling techniques are often assessed by MAE (Liu et al. 2020; Wang et al. 2020), we also use another comparison metric, average MAE, which is defined as

$$\overline{MAE}_{l,k,j} = \frac{\sum_{f,e} \left| \hat{y}_{kj}^{(f,l,e)} - y_{kj}^{(d)} \right|}{\sum_{f,e} 1}$$

for a lead time of l days. Taking

climatology as the reference forecast, we can define the MAE skill score for model m for each pixel as

$$MAE_SS_{l,k,j}^{(m)} = 1 - \frac{\overline{MAE}_{l,k,j}^{(m)}}{\overline{MAE}_{l,k,j}^{(clim)}} \tag{9}$$

Similarly, a higher MAE skill score is preferred.

For a fair comparison, skill scores presented in the following subsections exclude locations on the ocean as QM has results only on the Australian continent and Tasmania.

4.2 Results for forecasts made in 2012

Three typical ensemble members forecast on 1 Jan 2012 for 7 Jan 2012, downscaled by BI, VDSR and VDSO respectively, are illustrated in Fig. 1. VDSR keeps similar precipitation area patterns as BI-downscaled ACCESS-S1 forecasts and often has more areas with precipitation. Downscaled results of VDSO follow precipitation patterns of the raw forecasts and more likely reduce precipitation amount for 7 Jan 2012. VDSO could adjust the precipitation area shapes to some degree. For example, for ensemble members 2 and 11, VDSO substantially reduces the precipitation along the 30° S latitude line, which brings its downscaled precipitation images closer to the observations.⁵

Averaging across 48 initialisation days in 2012, we calculate its average CRPS skill score for each grid point for different lead times. Figure 4 illustrates average CRPS skill scores across the whole of Australia by the four downscaling models along with different lead times up to 216 days. VDSR has the highest scores in the first three lead times, and then VDSO becomes the best of the four models for almost all the other lead times. For example, for the lead time of 6 days (some typical downscaled results are illustrated in Fig. 1), its CRPS skill scores for the four models are spatially visualised in Fig. S1.⁶ For most locations on the Australian land (except north-western Australia, the eastern seaboard of Australia, and Tasmania), VDSO has a positive CRPS skill score. It has very high skills for locations in the central part of the Australian mainland where its three counterparts perform badly. The average CRPS skill score of VDSO is 5.69×10^{-2} . It is higher than 2.13×10^{-2} , -8.50×10^{-3} and -1.21×10^{-1} of VDSR, QM and BI respectively (the second column, Table 1). Averaging across the 217 different lead times, VDSO has positive CRPS skill scores for most locations in Australia, while its three counterparts have negative scores for most locations (Fig. S2). Their mean CRPS skill scores are 5.63×10^{-3} , -2.54×10^{-2} , -1.05×10^{-1} , and -1.42×10^{-1} , respectively. That means only VDSO is on average better than the climatology for most locations. Among the four downscaling techniques, only VDSO is better than climatology on average as its positive CRPS skill score.

To check the performance of these downscaling techniques for sub-seasonal forecasts, Fig. 5 illustrates the average CRPS skill scores for the first 45 lead times. The

skill scores of BI are around their mean of -1.39×10^{-1} for most locations in Australia. QM has some improvement with a mean of -7.40×10^{-2} . For most locations, VDSR has skill scores close to 0 with a mean of -4.65×10^{-3} . VDSO has positive skill scores for most locations on the Australian continent, with a mean of around 2.76×10^{-2} . VDSO still has negative skill scores along the eastern coastline, north-western parts of Australia, and Tasmania. VDSO has better forecast accuracy in terms of average MAE (Fig. S3). Except for the lead time of 0, VDSO has the lowest MAE values. Its average MAE is 1.37 mm/day for the first 45 lead times. VDSR comes second with an average MAE of 1.67 mm/day. They are much smaller than the other three methods, QM, BI and climatology, with the average MAE of 2.06, 2.29, and 2.20 mm/day respectively. The MAE skill scores of the four downscaling methods are 4.12×10^{-1} , 2.30×10^{-1} , 7.29×10^{-2} , and -9.90×10^{-2} respectively as listed in the third column, Table 1.

Similar to CRPS skill scores, the two deep learning methods hold their improvement for the long lead times in terms of the MAE skill scores as illustrated in Fig. 6. Except for the first six lead times, BI has negative MAE skill scores. QM often has positive MAE skill scores. Both deep learning models, VDSR and VDSO have substantial improvements for all the different lead times. Averaging across these 217 different lead times, the MAE skill scores of VDSO, VDSR, QM and BI are around 4.17×10^{-1} , 2.29×10^{-1} , 4.00×10^{-2} , and -9.48×10^{-2} respectively.

4.3 Results for forecasts made in 2010

Figure 7 illustrates the average CRPS skill scores along with different forecast lead times based on SCFs made on the 48 different initialisation days in 2010. For most of the 217 different lead times, VDSO has the highest CRPS skill score among the four models. For example, for the lead time of 6 days, the CRPS skill scores of two deep learning models are positive in most locations in Australia (Fig. S4). In comparison, QM and BI have negative CRPS skill scores in various locations. VDSO has higher CRPS skill scores than VDSR in the southeast and south-central Australia though both have quite similar spatial patterns. On average, the average CRPS skill scores of VDSO and VDSR are 5.34×10^{-2} and 3.97×10^{-2} . They are substantially higher than -3.01×10^{-2} and -8.79×10^{-2} of QM and BI respectively (Table 2). The average CRPS skill scores of VDSO and VDSR across the 217 different lead times are often positive or close to zero for most locations on Australian land (Figs. S5d and S5c), and both QM and BI are normally in the negative domain (Figs. S5b and S5a). The mean CRPS skill scores for VDSO, VDSR, QM

⁵ More downscaled rainfall images, as well as Python (v3.7.4) source codes of VDSO, can be found on <https://github.com/JiangWeiFanAI/HRSCF>.

⁶ To facilitate an easy comparison, these spatial plots use the same colour scale.

Fig. 4 Average CRPS Skill Scores across the Australian land for the ensemble forecasts made on the 48 different initialisation dates in 2012

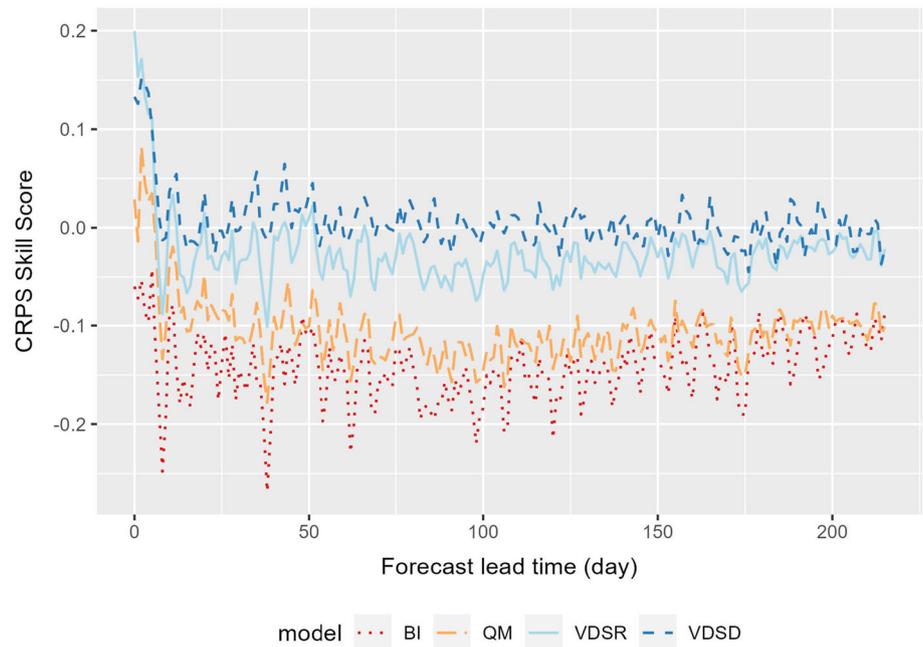


Table 1 Average CRPS and MAE (after ‘\’) skill scores of the four models on different lead times or periods for forecasts made on the 48 different initialisation dates in 2012

Model\Lead time	6-days only	Over 0–44 days	0–216 days
BI	− 1.21E−01\− 4.06E−02	− 1.39E−01\− 9.90E−02	− 1.42E−01\− 9.48E−02
QM	− 8.50E−03\1.54E−01	− 7.40E−02\7.29E−02	− 1.05E−01\4.00E−02
VDSR	2.13E−02\2.65E−01	− 4.65E−03\2.30E−01	− 2.54E−02\2.29E−01
VDSD	5.69E−02\4.11E−01	2.76E−02\4.12E−01	5.63E−03\4.17E−01
VDSD–BI	1.78E−01\4.52E−01	1.67E−01\5.12E−01	1.47E−01\5.12E−01

Higher skill scores are favoured

and BI across Australian land and 217 lead times are -1.02×10^{-2} , -2.53×10^{-2} , -6.46×10^{-2} , and -6.52×10^{-2} respectively (Table 2). VDSD is about 1.51×10^{-2} higher skill score than VDSR, and 5.50×10^{-2} higher than both the traditional downscaling techniques. VDSD is slightly worse than climatology on average for the 217 different lead times. Note that VDSD has 11, less than 22 in climatology, ensemble members that can lead to a few percentage points lower CRPS skill score (Ferro et al. 2008; Li and Jin 2020).

For sub-seasonal forecasts, the average CRPS skill scores for the first 45 different lead times are 1.38×10^{-2} , -1.02×10^{-3} , -6.62×10^{-2} , -9.06×10^{-2} , respectively, for VDSR, VDSR, QM and BI. As illustrated in Fig. 8, for most locations in Australia, both VDSR and VDSD have CRPS skill scores between -0.10 and 0.10 while VDSR has slightly higher CRPS skill scores in northern and eastern Australia. From the forecast accuracy, VDSR and VDSD have the average MAE values for the first 45 lead times around 2.14 and 2.07 mm/day respectively. They are smaller than 2.92, 3.16 and 2.74 mm/day obtained by QM,

BI, and climatology, respectively. Thus VDSR, VDSR, QM and BI again have MAE skill scores from high to low, as listed in Table 2.

The four downscaling techniques keep their MAE skill score orders almost for all the 217 different forecast lead times as illustrated in Fig. 9. VDSR and VDSR always have positive skill scores. Except for the first eight lead times, both QM and BI have negative skill scores. Averaging across these 217 lead times, the MAE skill scores of these four models are 2.41×10^{-1} , 2.13×10^{-1} , -9.14×10^{-2} and -1.97×10^{-1} respectively. VDSR has a relatively small improvement against VDSR, and both are much better than climatology. Considering these results, we conclude VDSR is comparable with climatology in terms of both forecast accuracy and ensemble forecast skill for the SCFs made on the 48 initialisation dates in 2010.

4.4 Discussions

From Figs. 4, 6, 7, and 9, we can see that VDSR normally has the highest average CRPS and MAE skill scores among

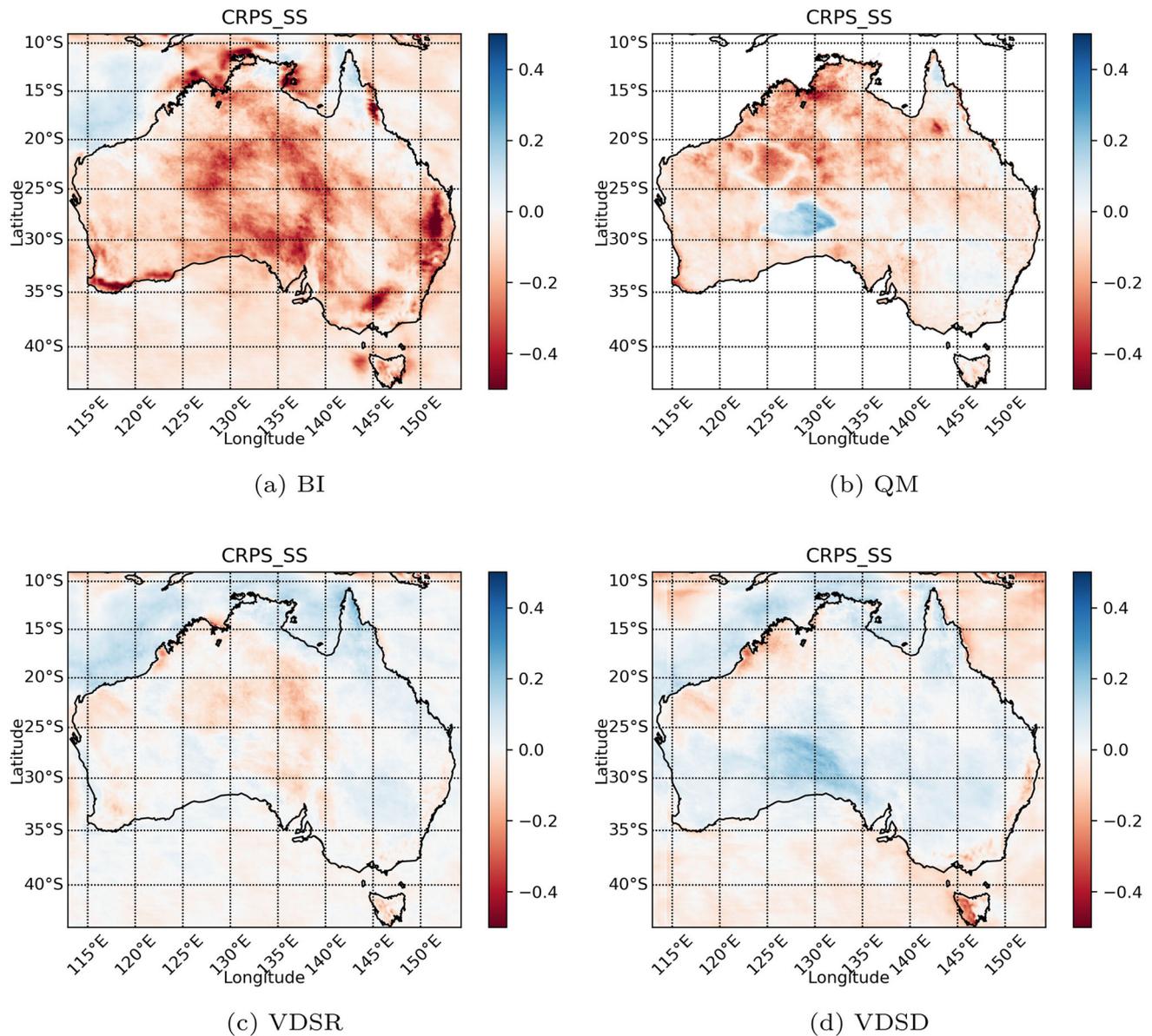


Fig. 5 Average CRPS skill score for lead time 0–44 days across Australia for forecasts made on the 48 different initialisation dates in 2012

the four downscaling models for different forecast lead times. The skills of ACCESS-S1 raw forecasts are often a bit away from climatology, indicated by the negative CRPS and the negative MAE skill scores of BI. VDSR often makes substantial improvements on downscaled forecasts. As listed in the last row of Table 1, VDSR has about 1.47×10^{-1} higher CRPS and 5.12×10^{-1} higher MAE skill scores than BI averaging across the whole of Australia and the 217 lead times on the SCFs made in 2012. The improvement can be slightly higher for the first 45 lead times. Similarly, for 2010 in Table 2, VDSR has about 5.50×10^{-2} higher CRPS skill scores and 4.38×10^{-1} higher MAE skill scores on average. Thus, the bias and displacement issues in the ACCESS-S1 raw forecasts are

mitigated to some degree. VDSR also makes substantial improvements over BI, though it performs slightly worse performance than VDSR. The improvement of VDSR and VDSR may come from their large receptive fields of around $492 \text{ km} \times 492 \text{ km}$ due to the 20 convolutional layers, which could benefit local forecasts from GCM's forecast skills on a large scale. The relatively small image patches used for training the other two outstanding SISR models, RCAN and ESRGAN, would fail to capture GCM's forecast skills on a large scale. The further improvement of VDSR over VDSR illustrates the usefulness of incorporating other climate variables into downscaling.

Fig. 6 Average MAE skill scores across Australia for daily precipitation forecasts made on the 48 different initialisation dates in 2012

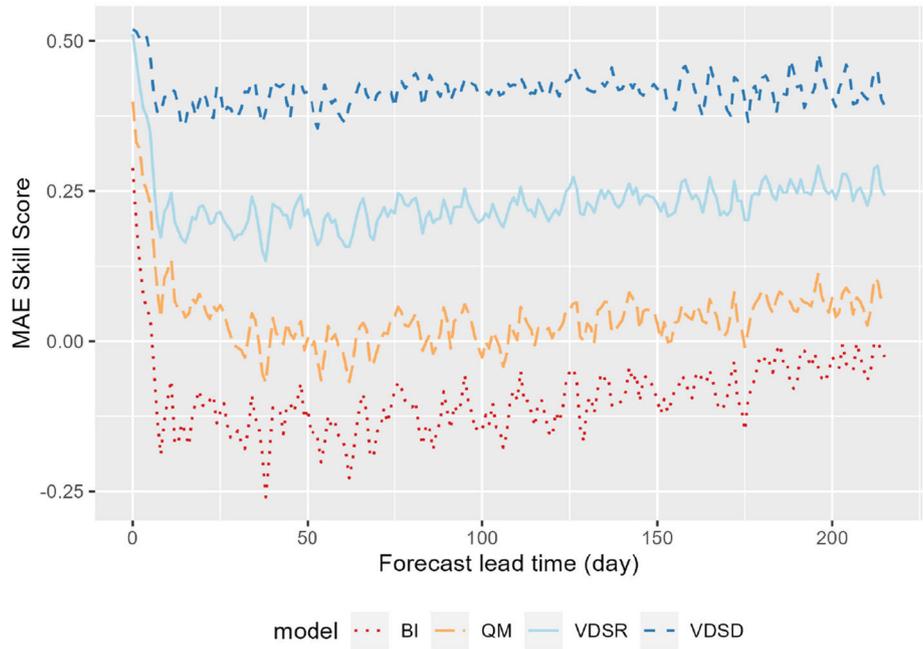


Fig. 7 Average CRPS skill scores across the Australian land for forecasts made on the 48 initialisation dates in 2010

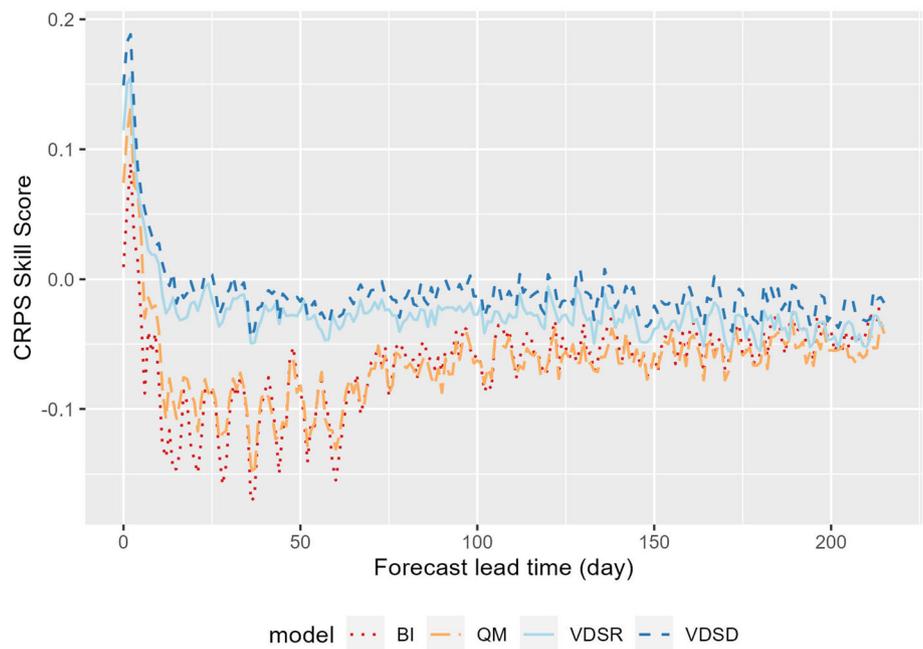


Table 2 Average CRPS and MAE (the higher the better) of the four models on different lead times or periods for forecasts made on the 48 different initialisation dates in 2010

Model\Lead time	6 days only	Over 0–44 days	0–216 days
BI	− 8.79E−02\− 1.68E−01	− 9.06E−02\− 2.32E−01	− 6.52E−02\− 1.97E−01
QM	− 3.01E−02\6.38E−02	− 6.62E−02\− 8.37E−02	− 6.46E−02\− 9.14E−02
VDSR	3.97E−02\3.11E−01	− 1.02E−03\2.36E−01	− 2.53E−02\2.13E−01
VDSD	5.34E−02\3.42E−01	1.38E−02\2.71E−01	− 1.02E−02\2.41E−01
VDSD–BI	1.41E−01\5.10E−01	1.04E−01\5.02E−01	5.50E−02\4.38E−01

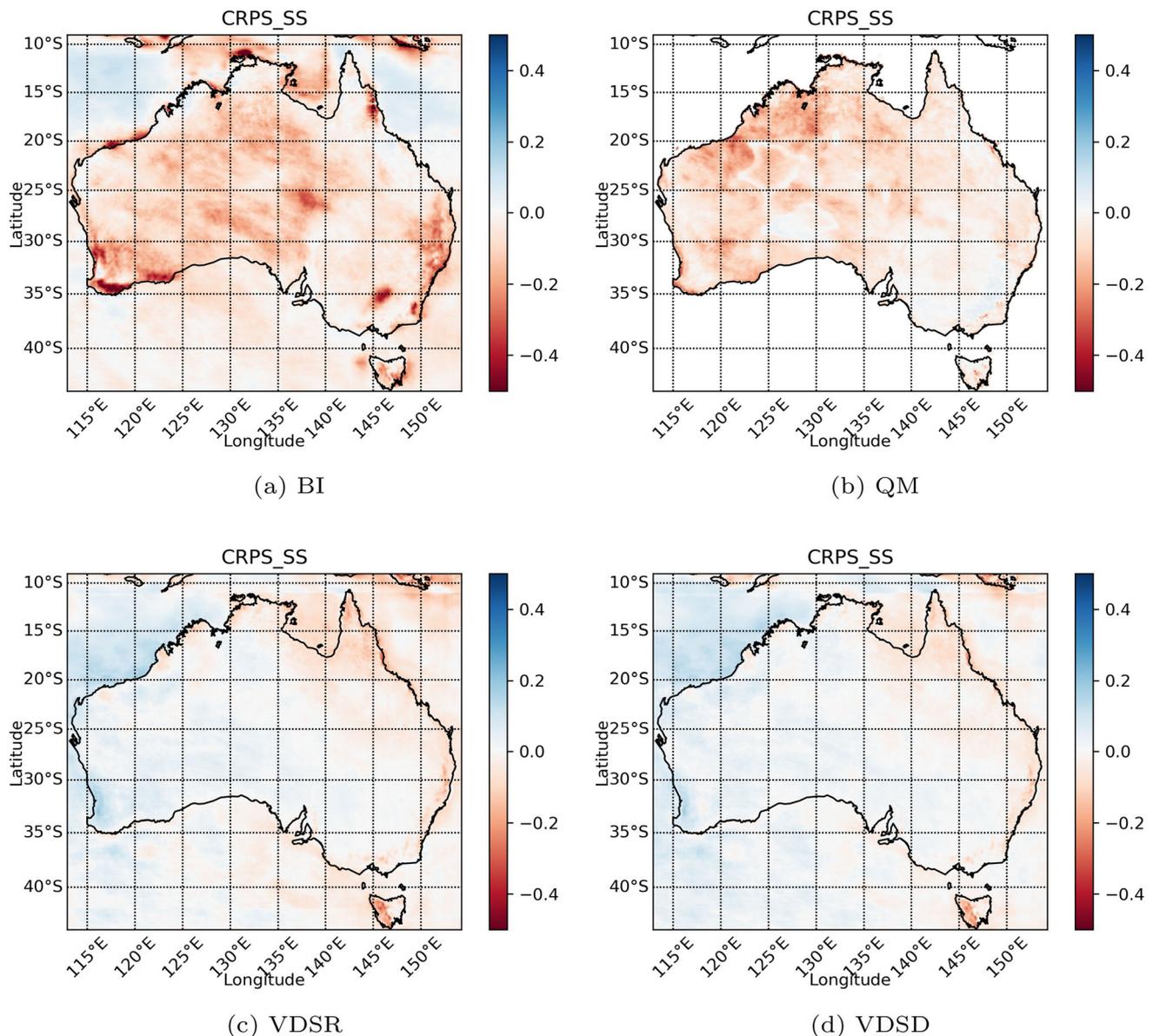


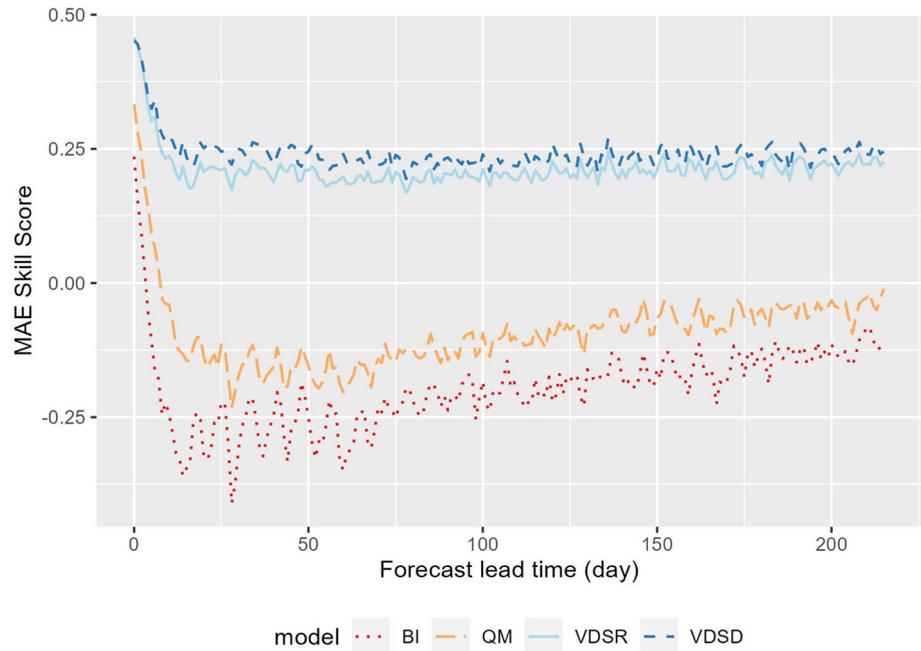
Fig. 8 Average CRPS skill score for lead time 0–44 days across Australia for forecasts made in 2010

Though VDSD makes a substantial improvement, for the SCFs made in 2010, its average overall ensemble forecast skill, CRPS skill score, is slightly worse than climatology in general. There are some possible reasons. (1) For forecasts made in 2010, rainfall data from 1 Jan 2010 to 29 July 2011 are used for skill assessment. The years 2010 and 2011 are the third- and second-wettest calendar years on record for Australia, with 703 mm and 708 mm respectively. Both are well above the long-term average of 465 mm due to the La Niña event peak.⁷ The La Niña event peak in 2012 is much weaker and made 2012

relatively easier to forecast. That means the training data for the models to test on 2010 and 2011 have relatively less precipitation, hence VDSD intends to move in that direction, which deteriorated its performance for the SCFs made in 2010. (2) The climatology benchmark we have used has 22 ensemble members. Such a double ensemble size can lead to a few percentage points higher CRPS skill scores (Ferro et al. 2008). (3) The host climate model ACCESS-S1 may perform worse in 2010 than in 2012, on e.g., geopotential heights. For both test settings, VDSD's final performance still heavily depends on ACCESS-S1 raw forecasts. Therefore, to generate skilful SCFs, more research efforts should be put into improving both deep-

⁷ <http://www.bom.gov.au/climate/history/enso/>. Accessed on 20 Jan 2022.

Fig. 9 Average MAE skill scores across Australia for precipitation forecasts made on the 48 initialisation dates in 2010



learning downscaling methods like VDS D and climate models like ACCESS-S1.

To compare other downscaling techniques for Australia, the deep learning techniques are preliminarily examined against station-based techniques: Extended Copula-based Post-Processing (E CPP) (Li and Jin 2020), and an improved analogue method (Shao and Li 2013), as well as dynamic downscaling CCAM (Thatcher and McGregor 2009). The analogue method (Shao and Li 2013) is a specific implementation of the well-known K-Nearest Neighbours (KNN) method in the machine learning community. As most station-based downscaling techniques are time-consuming to train models (like E CPP) or to make forecasts (like analogue) for the whole of Australia, we only choose five weather stations for a simple comparison. These five stations are from three different states in Australia and have different climate characteristics (Table S4). The forecast lead time is up to 28 days as the dynamic downscaling like CCAM generates a huge amount of data for its 11 runs with different boundary conditions from 11 ensemble members for each SCF initialisation date. Similarly, the initiation date is restricted to the first day of each month in 2012. The CCAM downscaling used in this comparison employed a better soil initialisation with spin-up soil temperature and moisture, which have improved forecast skills. The average CRPS skill score for a single lead time is averaged from the five locations and 12 different initialisation days. For the first week (with a lead time of up to 6 days), all the downscaling techniques except BI have a positive skill score (the second column in Table S5). E CPP performs best, followed by two deep

learning downscaling techniques VDS D and VDS R. For the first 2 weeks (lead time up to 13 days), VDS D performs best on average, followed by VDS R, E CPP and QM. Averaging over 29 different lead times, VDS D is better than E CPP, QM and VDS R that have similar forecast skills as climatology. Analogue, BI and CCAM have negative skill scores as the raw seasonal forecasts lose skills with a longer lead time for these five weather stations. VDS D often has better forecast skills than its counterparts for the three different lead time periods. These improvements are relatively small compared with the large standard deviation of CRPS skill scores (Table S5), though we can claim at least that VDS D has comparable forecast skills as its competitors.

Another comparison can be made from the perspective of computational time. Table 3 lists the average computation time required for both training and operation/test where 11 ensemble members for 217 days forecasts from ACCESS-S1 are downscaled. We used the following hyper-parameters in both VDS R and VDS D. The number of epochs was 50, for which we normally observed the

Table 3 Computation time in hours of four downscaling methods

Method	Training time on Gadi	Operation time on the PC
BI	0	0.02
QM	0	11.21
VDS R	12.12	0.08
VDS D	16.76	0.56

objective function stabilised. The learning rate was 1.0×10^{-4} , relatively small as both networks are very deep (Bengio et al. 1994). The optimisation method was stochastic gradient descent with a momentum of 0.9. The deep learning training was run on Gadi, a high-performance computer in National Computational Infrastructure (NCI), Australia. We used 36 CPUs (Intel®Xeon® Platinum 8268, 2.9GHz), and 3 GPUs (Nvidia® V100). Forecast downscaling and validation were done on a normal PC with one Intel®Core™ i5-9600K processor (3.7GHz) and with a mid-range GPU GeForce RTX 2070. The training time for VDSD was around 16.76 h. It is about 38.3% longer than VDSR. BI and QM didn't require training time. Downscaling operation on the PC, BI, QM, VDSR and VDSD required 0.02, 11.21, 0.08 and 0.56 h respectively. VDSD is about 7 times slower than VDSR and 20 times faster than QM.

ECPP (Li and Jin 2020) needed about 1 h for training for a station or grid point and took 0.46 s for an operation forecast. It would take around 15.11 h to downscale rainfall for the whole of Australia on a normal PC. The improved analogue method took similar time as ECPP for downscaling but didn't need training (after an appropriate similarity metric was specified). CCAM (Thatcher and McGregor 2009), a dynamic downscaling model, doesn't need training, and took about 0.33 h to simulate a single 1-month lead time forecast to 10 km resolution on a CSIRO supercomputer Pearcey with 1536 cores (personal communication with M. Thatcher). Compared with ECPP and CCAM, VDSD is much faster for downscaling long lead time daily forecasts.

5 Conclusion

Downscaling long lead time daily rainfall ensemble forecasts has been modelled as a single image super-resolution (SISR) problem with an additional target on maximising overall ensemble forecast skill—continuous ranked probability score (CRPS). To leverage advanced deep learning techniques, we have applied three outstanding SISR techniques to generate high-level features and learn non-linear relationships automatically, and chosen very deep super-resolution (VDSR) as the most suitable model. The selection has been based on average CRPS on a randomly selected validation data set. We have incorporated an extra climate variable, geopotential height, into VDSR and established the very deep statistical downscaling (VDSD) model with an expectation to enhance downscaling. Both deep learning models have finalised their structures based on the average CRPS on the validation data. On leave-one-year-out cross-validation for 48 ensemble SCFs made in

2012 and 2010, VDSD has outperformed VDSR and two traditional downscaling techniques QM and BI in terms of both forecast accuracy and skill on the whole of Australia. With positive CRPS skill scores in general in Australia, VDSD has outperformed climatology, a benchmark for long lead time ensemble climate forecast, in 2012 and the first 45 lead times in 2010, though its improvement has become smaller with longer forecast lead time. As evidenced by its forecast skill improvement over the ACCESS-S1 forecasts in 2012 and 2010, optimising CRPS in our model development might have mitigated the mismatch issues between the ACCESS-S1 low-resolution raw forecasts and high-resolution observations to some degree. A simple comparison with two station-based statistical downscaling methods, an improved analogue method and ECPP, and one dynamic downscaling model CCAM on five representative weather stations has further demonstrated VDSD's advantages. Compared with the CRPS skill score variabilities, the improvement of VDSD over VDSR or ECPP is relatively small. VDSD is comparable with, if not better than, various downscaling counterparts in terms of forecast skill and accuracy. On the other hand, both VDSD and VDSR, after lengthy training over a large amount of data, have downscaled long lead time daily precipitation very fast. They have been normally much faster than sophisticated station-based downscaling techniques like ECPP and analogue, and dynamic downscaling like CCAM for the whole of Australia. Thus, deep learning models, such as the proposed VDSD, have demonstrated their potential for possible operational use in the future.

Though deep-learning-based downscaling methods can provide more skilful high-resolution SCFs to drive impact models or biophysical models, the accuracy and overall forecast skills of these SCFs may still not be high enough for direct use by wider communities such as agriculture and hydrology (Kusunose and Mahmood 2016; Luo 2016). There are several directions to move the proposed technique for daily operation in the future. Station-based precipitation observations have not been assimilated in BARRA and its grid precipitation may not be very consistent with on-the-ground observations (Acharya et al. 2019). To remove such inconsistency, station-based downscaling techniques might further improve long lead time forecasts. Integrating station-based techniques with deep learning and conduct comprehensive comparison are subject to future work. VDSD only downscales to 12 km, which should be further enhanced to a higher resolution for real-world applications, by including other inputs (Vandal et al. 2017; Pan et al. 2019). Deep-learning-based downscaling still requires a lot of time for model development, structure finalisation, and parameter training, and we leave downscaling techniques for other climate variables as future work. For a fair comparison (and saving training

time), we have only included the forecasts with forecast lead times less than 7 days into the training data. It would lead our models to take less attention to correcting inherent biases of GCM's long lead time forecasts. Skilful SCFs depend more heavily on progresses made in climate modelling. Thus, to deliver skilful SCFs to final applications, climate modelling and deep learning communities should collaborate closely for further development.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00477-023-02444-x>.

Acknowledgements The authors would like to thank the reviewers and the associate editor for their valuable comments and thoughtful suggestions, which have helped us to improve the manuscript. The authors would like to thank multiple CSIRO colleagues and BoM researchers, including Andrew Moore, Marcus Thatcher, Andrew Powell, Yanchang Zhao, Robert Smalley, and Morwenna Griffiths for their discussion and helps in this work. This research was undertaken with the assistance of resources and services from the National Computational Infrastructure (NCI), which is supported by the Australian Government.

Author contributions All authors contributed to the study's conception and design. Material preparation, data collection and analyses were performed by HJ, WJ, and MC. The first draft of the manuscript was mainly prepared by HJ and WJ and all the authors commented on previous versions of the manuscript. All the authors read and approved the final manuscript.

Funding Open access funding provided by CSIRO Library Services. This work was partially funded by the CSIRO Digiscape Future Science Platform and the CAS-CSIRO Partnership program.

Data availability All the data sets used in this study are stored on NCI.org.au, which is accessible after applying for the following research project membership. ACCESS-S1 raw forecast data and their calibrated versions are stored under the project "Seasonal Prediction ACCESS-S1 Hindcast" (g/data/ub7/), BARRA reanalysis data are under the project of "Australian Regional Reanalysis" (g/data/ma05/). The intermediate results of this study are stored on /scratch/iu60/wj1671/ under the project "High Resolution Seasonal Climate Forecast". Python source codes, some downscaled images and results are available in the repository, <https://github.com/JiangWeiFanAI/HRSCF>.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright

holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- Acharya SC, Nathan R, Wang QJ et al (2019) An evaluation of daily precipitation from a regional atmospheric reanalysis over Australia. *Hydrol Earth Syst Sci* 23(8):3387–3403. <https://doi.org/10.5194/hess-23-3387-2019>
- Ahmadalipour A, Moradkhani H, Rana A (2018) Accounting for downscaling and model uncertainty in fine-resolution seasonal climate projections over the Columbia river basin. *Clim Dyn* 50(1–2):717–733. <https://doi.org/10.1007/s00382-017-3639-4>
- Baño-Medina J, Manzanar R, Gutiérrez JM (2020) Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geosci Model Dev* 13(4):2109–2124
- Basso B, Liu L (2019) Seasonal crop yield forecast: methods, applications, and accuracies, vol 154. Elsevier, Amsterdam, pp 201–255. <https://doi.org/10.1016/bs.agron.2018.11.002>
- Bengio Y, Simard P, Frasconi P (1994) Learning long-term dependencies with gradient descent is difficult. *IEEE Trans Neural Netw* 5(2):157–166
- Bettolli M, Solman S, Da Rocha R et al (2021) The CORDEX flagship pilot study in southeastern south America: a comparative study of statistical and dynamical downscaling models in simulating daily extreme precipitation events. *Clim Dyn* 56(5):1589–1608
- Bureau National Operations Centre (2019) Operational implementation of ACCESS-S1 forecast post processing. Tech. Rep. 124, Bureau of Meteorology, Melbourne VIC 3001
- Crimp S, Jin HD, Kocic P et al (2019) Possible future changes in south east Australian frost frequency: an inter-comparison of statistical downscaling approaches. *Clim Dyn* 52(1–2):1247–1262. <https://doi.org/10.1007/s00382-018-4188-1>
- Dong C, Loy CC, He K et al (2014) Learning a deep convolutional network for image super-resolution. In: European conference on computer vision. Springer, pp 184–199
- Espoholt L, Agrawal S, Sønderby C et al (2022) Deep learning for twelve hour precipitation forecasts. *Nat Commun* 13(1):1–10. <https://doi.org/10.1038/s41467-022-32483-x>
- Ferro CA, Richardson DS, Weigel AP (2008) On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorol Appl* 15(1):19–24
- Grimm EP, Gneiting T, Berrocal VJ et al (2006) The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Q J R Meteorol Soc* 132(621C):2925–2942
- Hersbach H (2000) Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather Forecast* 15(5):559–570. [https://doi.org/10.1175/1520-0434\(2000\)015<0559:DOTCRP>2.0.CO;2](https://doi.org/10.1175/1520-0434(2000)015<0559:DOTCRP>2.0.CO;2)
- Hudson D, Alves O, Hendon HH et al (2017) ACCESS-S1 the new bureau of meteorology multi-week to seasonal prediction system. *J South Hemisphere Earth Syst Sci* 67(3):132–159. <https://doi.org/10.1071/ES17009>
- Jin H, Li M, Hopwood G et al (2022) Improving early-season wheat yield forecasts driven by probabilistic seasonal climate forecasts. *Agric For Meteorol* 315(108):832. <https://doi.org/10.1016/j.agrformet.2022.108832>
- Johnson SJ, Stockdale TN, Ferranti L et al (2019) SEAS5: the new ECMWF seasonal forecast system. *Geosci Model Dev* 12(3):1087–1117. <https://doi.org/10.5194/gmd-12-1087-2019>
- Kim J, Kwon Lee J, Mu Lee K (2016) Accurate image super-resolution using very deep convolutional networks. In:

- Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1646–1654
- Kusunose Y, Mahmood R (2016) Imperfect forecasts and decision making in agriculture. *Agric Syst* 146:103–110. <https://doi.org/10.1016/j.agsy.2016.04.006>
- Ledig C, Theis L, Huszár F et al (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4681–4690
- Li M, Jin H (2020) Development of a postprocessing system of daily rainfall forecasts for seasonal crop prediction in Australia. *Theor Appl Climatol* 141:1331–1349. <https://doi.org/10.1007/s00704-020-03268-3>
- Liu Y, Ganguly AR, Dy J (2020) Climate downscaling using YNet: a deep convolutional network with skip connections and fusion. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery and data mining, pp 3145–3153
- Luo Q (2016) Necessity for post-processing dynamically downscaled climate projections for impact and adaptation studies. *Stoch Environ Res Risk Assess* 30(7):1835–1850. <https://doi.org/10.1007/s00477-016-1233-7>
- Manzanas R (2020) Assessment of model drifts in seasonal forecasting: sensitivity to ensemble size and implications for bias correction. *J Adv Model Earth Syst* 12(3):e2019MS001751
- Maraun D, Widmann M (2018) Statistical downscaling and bias correction for climate research. Cambridge University Press, Cambridge
- Merryfield WJ, Baehr J, Batté L et al (2020) Current and emerging developments in subseasonal to decadal prediction. *Bull Am Meteorol Soc* 101(6):E869–E896
- Michelangeli PA, Vrac M, Loukos H (2009) Probabilistic downscaling approaches: application to wind cumulative distribution functions. *Geophys Res Lett.* <https://doi.org/10.1029/2009gl038401>
- Pan B, Hsu K, AghaKouchak A et al (2019) Improving precipitation estimation using convolutional neural network. *Water Resour Res* 55(3):2301–2321. <https://doi.org/10.1029/2018WR024090>
- Rad MS, Bozorgtabar B, Marti UV et al (2019) SROBB: targeted perceptual loss for single image super-resolution. In: Proceedings of the IEEE/CVF international conference on computer vision, pp 2710–2719
- Ratnam J, Dai T, Behera SK (2017) Dynamical downscaling of SINTEX-F2v CGCM seasonal retrospective austral summer forecasts over Australia. *J Clim* 30(9):3219–3235
- Reichstein M, Camps-Valls G, Stevens B et al (2019) Deep learning and process understanding for data-driven earth system science. *Nature* 566(7743):195–204
- Rodrigues ER, Oliveira I, Cunha R et al (2018) DeepDownscale: a deep learning strategy for high-resolution weather forecast. In: 2018 IEEE 14th international conference on e-science (e-science), pp 415–422
- Saha S, Moorthi S, Wu X et al (2014) The NCEP climate forecast system version 2. *J Clim* 27(6):2185–2208. <https://doi.org/10.1175/JCLI-D-19-0230.1>
- Şan M, Nacar S, Kankal M et al (2022) Daily precipitation performances of regression-based statistical downscaling models in a basin with mountain and semi-arid climates. *Stoch Environ Res Risk Assess.* <https://doi.org/10.1007/s00477-022-02345-5>
- Schepen A, Everingham Y, Wang QJ (2020) An improved workflow for calibration and downscaling of GCM climate forecasts for agricultural applications—a case study on prediction of sugarcane yield in Australia. *Agric For Meteorol* 291(107):991. <https://doi.org/10.1016/j.agrformet.2020.107991>
- Shao Q, Li M (2013) An improved statistical analogue downscaling procedure for seasonal precipitation forecast. *Stoch Environ Res Risk Assess* 27(4):819–830. <https://doi.org/10.1007/s00477-012-0610-0>
- Shi X, Chen Z, Wang H et al (2015) Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems, pp 802–810
- Su CH, Eizenberg N, Steinle P et al (2019) BARRA v1.0: the bureau of meteorology atmospheric high-resolution regional reanalysis for Australia. *Geosci Model Dev* 12(5):2049–2068. <https://doi.org/10.5194/gmd-12-2049-2019>
- Thatcher M, McGregor JL (2009) Using a scale-selective filter for dynamical downscaling with the conformal cubic atmospheric model. *Mon Weather Rev* 137(6):1742–1752
- The Centre for International Economics (2014) Analysis of the benefits of improved seasonal climate forecasting for agriculture. Tech. rep., Managing Climate Variability Program. <http://www.climatekelpie.com.au/Files/MCV-CIE-report-Value-of-improved-forecasts-non-agriculture-2014.pdf>. Accessed Nov 2020
- Vandal T, Kodra E, Ganguly S et al (2017) DeepSD: generating high resolution climate change projections through single image super-resolution. In: KDD'17, pp 1663–1672. <https://doi.org/10.1145/3097983.3098004>
- Wang X, Yu K, Wu S et al (2018) ESRGAN: enhanced super-resolution generative adversarial networks. In: Proceedings of the European conference on computer vision (ECCV) workshops, pp 63–79
- Wang Z, Chen J, Hoi SC (2020) Deep learning for image super-resolution: a survey. *IEEE Trans Pattern Anal Mach Intell* 43(10):3365–3387
- Wang F, Tian D, Lowe L et al (2021) Deep learning for daily precipitation and temperature downscaling. *Water Resour Res* 57(4):e2020WR029308
- Zhang Y, Li K, Li K et al (2018a) Image super-resolution using very deep residual channel attention networks. In: Proceedings of the European conference on computer vision (ECCV), pp 286–301
- Zhang Y, Tian Y, Kong Y et al (2018b) Residual dense network for image super-resolution. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2472–2481

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Authors and Affiliations

Huidong Jin¹  · Weifan Jiang^{1,2} · Minzhe Chen² · Ming Li³  · K. Shuvo Bakar^{1,4}  · Quanxi Shao³ 

✉ Huidong Jin
warren.jin@csiro.au

Weifan Jiang
u6683698@alumni.anu.edu.au

Minzhe Chen
chenminzhe2020@163.com

Ming Li
Ming.Li@csiro.au

K. Shuvo Bakar
shuvo.bakar@sydney.edu.au

Quanxi Shao
Quanxi.Shao@csiro.au

¹ Data61, CSIRO, North Science Road, Acton, ACT 2601, Australia

² CECS, The Australian National University, North Road, Acton, ACT 2601, Australia

³ Data61, CSIRO, P.O. Box 1130, Bentley, WA 6102, Australia

⁴ Faculty of Medicine and Health, University of Sydney, Science Rd, Camperdown, NSW 2050, Australia