Long Lead Time Probabilistic Daily Rainfall Forecasts through Deep Learning for Australia

Huidong Jin[*], *Weifan Jiang*, *Minzhe Chen*, *To be finalised*

Abstract

Skilful and high-resolution daily weather forecasts for upcoming seasons are of huge value to climate-sensitive sectors, especially for agriculture and construction sectors. General Circulation Models (GCM) are routinely providing long lead time ensemble climate forecasts while requiring downscaling techniques to improve their spatial resolution and consistency with local observations to be accurate and skilful. Traditional downscaling techniques, which use historical climate data to learn some relationship from low-resolution to finer-resolution, have more or less skill improvement but are often time-consuming or labour-intensive for operation for big regions. While downscaling techniques based on image super-resolution have successfully been developed, almost all of them focused on simplified situations where low resolution images match really well with high-resolution images, which are not the case in long lead time daily rainfall forecasts. In this paper, after applying several deep learning models for downscaling problems, we choose Very Deep Super-Resolution (VDSR) as the most suitable candidate, according to an overall skill metric, Continuous Ranked Probability Score (CRPS) that considers both accuracy and uncertainty of ensemble forecasts. We then propose Very Deep Statistical Downscaling (VDSD) model via incorporating resolved climate variables such as geopotential height, from which extra features are extracted to improve forecast performance. Both VDSR and VDSD are tested on downscaling ACCESS-S1 60km rainfall forecasts to 12km BARRA rainfall data with up to 216 days lead time for Australia. Leave-one-year-out cross-validation results illustrate that VDSD has higher forecast accuracy and skill, measured by Mean Absolute

[*]Corresponding author. Email: Warren.Jin@csiro.au

Error (MAE) and CRPS respectively, than VDSR and the standard downscaling techniques including quantile mapping. The results also show that VDSD performs better than or similarly to climatology, a benchmark for long lead time climate forecast. VDSD is lightweight to run on a modest computer system. After further development, this deep learning model can be used for downscaling operations.

Keywords: Statistical downscaling, probabilistic forecast, ensemble forecasts, image super-resolution, seasonal climate forecasts

1 Introduction

Seasonal Climate Forecasts (SCF) have great value to multiple socioeconomic sectors such as agriculture, construction, mining, tourism, energy, and health (Merryfield et al., 2020; Manzanas, 2020). For example, daily rainfall forecasts for upcoming seasons can benefit the whole agriculture value chain, such as helping farmers adapt their farm planning and management, and insurers and traders adjust their pricing scheme. SCFs have been estimated to contribute between $0 to $8.40 per hectare per year for USA agriculture, between -A$21 to A$258 per hectare per year in Australia, based on a large number of studies (Parton et al., 2019; Mjelde and Griffiths, 1998). For the the whole of Australia, potential annual value added from skilful SCFs would be around A$1.6 billion for the agricultural sector and A$192 million for the construction sector (The Centre for International Economics, 2014). As climate change increase both the variability and uncertainty of weather patterns, the value of SCFs would also further increase (Kusunose and Mahmood, 2016). To realise their full potentials, SCFs provided should be timely and skilful in high spatial resolution so as to help these weather-sensitive sectors make evidence-based site-specific decisions (Li and Jin, 2020; Schepen et al., 2020).

After the last three decades of development, SCFs using General Circulation Models (GCMs) have moved beyond the research realm and are routinely produced by climate forecast centres around the world (Hudson et al., 2017; Merryfield et al., 2020; Saha et al., 2014; Johnson et al., 2019). These state-of-the-art GCMs couple together physics-based models of ocean, atmosphere, land surface and sea-ice. They can capture synoptic-scale climate dynamics. GCMs are gridded with horizontal/spatial resolutions commonly

around 100 km (Vandal et al., 2017; Ratnam et al., 2017; Liu et al., 2020; Johnson et al., 2019). These physical models also incorporate hundreds of semi-empirical relationship to approximate processes such as convection and cloud formation that are too fine for the models to resolve (Vandal et al., 2017; Manzanas, 2020). These empirical relationships may be ill-constrained. Limited by computational resources, coarse spatial resolution and simplified nature of GCMs often lead them to produce forecasts not reliably consistent with observed weather, especially for precipitation or with longer lead time. To improve forecast skills and quantifying uncertainty, ensemble forecasts, i.e., multiple simulations of a single model each with different initial conditions and/or parameters, are normally carried out and published (Merryfield et al., 2020). For example, the operational SCFs from Australia's Bureau of Meteorology (BoM) have 11 ensemble members for each initialisation date and its United States' counterpart 40 members (Saha et al., 2014). The coarse spatial resolution and low forecast quality in representing local climate characteristics of GCMs circumvent their applications in weather-sensitive sectors (Baño-Medina et al., 2020; Schepen et al., 2020; Kusunose and Mahmood, 2016). The barriers may be bridged via downscaling techniques which generate more skilful and localised forecasts by making use of weather observations, and sometimes other localised information (Maraun and Widmann, 2018; Bettolli et al., 2021).

Downscaling is generally difficult and computationally expensive because of the complex nature of the spatial-temporal structure of high-resolution climate variables, especially for precipitation. There is a large body of downscaling techniques developed, including dynamical downscaling (Ratnam et al., 2017; Thatcher and McGregor, 2009; Manzanas, 2020), statistical downscaling (Maraun and Widmann, 2018), and recent development on deep learning-based downscaling. Comparisons between traditional statistical and dynamical downscaling suggest that neither group of methods are clearly superior, however in practice computationally cheaper statistical methods are widely used (Baño-Medina et al., 2020). The skill improvement of statistical downscaling for long lead time daily forecasts can be substantial or almost nothing, depending on locations and seasons (Schepen et al., 2020; Li and Jin, 2020; Manzanas et al., 2018). The inadequacy of these statistical downscaling techniques may stem from the pre-engineered features and relationships before the modelling process, rarely exploiting their spatio-temporal dependencies exhaustively. This limits their ability to capture important information beyond prior knowledge (Baño-Medina et al., 2020; Liu et al., 2020). Auto-

matic feature extraction and selection integrated into the modelling process with deep learning, especially convolutional neural networks (CNNs), has achieved notable success in modelling data with spatial context, recently on climate science (Reichstein et al., 2019). Deep learning has successfully been used in precipitation nowcasting (Shi et al., 2017; Luo et al., 2021; Xingjian et al., 2015), which predict rainfall intensity in a local region over a relatively short period of time, and precipitation parameterisations from GCMs (Pan et al., 2019). More related to our objective in this paper, several downscaling techniques have been developed based on Single Image Super-Resolution (SISR) techniques since 2017 (Vandal et al., 2017). To reuse deep learning based SISR techniques, most works over-simplify downscaling problems by taking the upscaled observations or finer-resolution reanalysis data, rather than raw GCMs out, as the input, such that the low-resolution input images match really well with their corresponding high-resolution outputs, and bias-correction required by long lead time forecasts is left behind (Vandal et al., 2017; Rodrigues et al., 2018; Liu et al., 2020; Wang et al., 2021). Some works use relatively shallow convolution layers and their downscaling performance may not be better than classic downscaling approaches (Baño-Medina et al., 2020; Pan et al., 2019; Wang et al., 2021).

Statistical downscaling looks similar to SISR as both aim at getting higher resolution images from lower resolution images if climate variable data are treated as images (Liu et al., 2020). However, there are several differences.

1. Inputs and outputs in downscaling are clearly from different sources, such as low-resolution forecasts from GCM vs historical weather data (Liu et al., 2020). In SISR, the low-resolution input images and high-resolution target images are arguably from the same source, e.g., the high-resolution images are often aggregated to form low resolution images as the inputs (Wang et al., 2020). Most deep learning based downscaling techniques focused on a single source (Vandal et al., 2017; Wang et al., 2021).

2. Statistical downscaling often use additional 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 and topographical information (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).

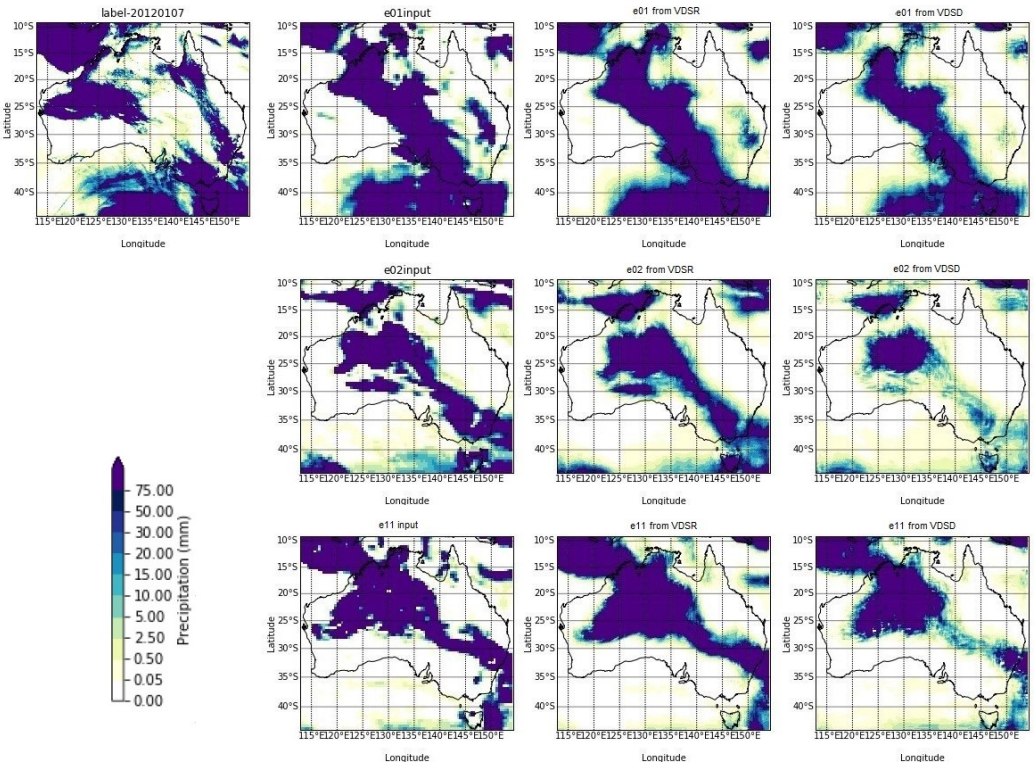


Fig. 1: Label image and ensemble rainfall forecasts for 7 Jan 2012 with six-day lead time for the forecasts made on 1 Jan 2012. Images in the four columns are the high-resolution label image from BARRA, ensemble member forecasts from ACCESS-S1 after bicubic interpolation, downscaled results of VDSR and VDSD respectively. Only the first, the second, the 11th members are illustrated.

3. Bias and displacement in space or time are common in climate forecasts, especially for precipitation, due to the inherent complexity in climate modelling. To mitigate these issues, multiple possible forecast trajectories are provided as a practice 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 between two images and overall forecast skill as ensemble forecasts by considering forecast uncertainty (Grimit et al., 2006; Li and Jin, 2020; Kusunose and Mahmood, 2016). The latter is predominant in climate communities (Grimit et al., 2006; Ferro et al., 2008; Schepen et al., 2020) but, as far as we know, has never been used in deep learning downscaling development.

To address these differences, we choose Very Deep Super-Resolution (VDSR) (Kim et al., 2016) from several SISR techniques as a suitable candidate for our downscaling problem 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). Precipitation raw 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 finalised based on CRPS on a randomly selected validation data subset. It is tested on real-world application scenarios. 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, measured by Mean Absolute Error (MAE) and CRPS respectively. In addition, its performance is better than or comparable with climatology, a benchmark for long lead time climate forecasts.

In the remaining of this paper, we briefly discuss related works in Section 2 and present climate data in Section 3. We select three SISR models and propose and finalise the new downscaling model VDSD in Section 4. Cross-validation results and comparison are given in Section 5. We conclude the paper in Section 6 with discussions on further developments of VDSD for possible operational use.

2 Related works

2.1 Image Super-Resolution

Single Image Super-Resolution (SISR) is the recovery of a high-resolution image from a low-resolution one. The low resolution image L is often regarded as the result of degradation $L = \mathcal{D}(H; \gamma)$ where D is degradation mapping function; H is high-resolution image corresponding to L; γ is the parameters of the degradation mapping function (Wang et al., 2020). Various super-resolution data sets are actually obtained by various aggregation or degradation mapping. A series of low- and high-resolution image pairs have been created, and researchers would like to generate high solution images from low-resolution ones:

$$S = \mathcal{F}(L; \theta) \tag{1}$$

where $\mathcal{F}$ is the super-resolution mapping function and θ is its parameter. All SISR works are to locate a suitable function $\mathcal{G}$ and its parameter θ.

The simplest SISR techniques are spatial interpolation, such as nearest-neighbour interpolation, bilinear interpolation, and bicubic interpolation. Bilinear interpolation is performed using linear interpolation first in one direction from one resolution to another resolution, and then again in the other direction. Bicubic Interpolation (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 artifacts. BI is often considered as the baseline for spatial downscaling of precipitation fields (Vandal et al., 2017).

Since the seminal work by Dong et al. (2014), deep learning based SISR techniques have been widely developed and successfully applied. Most of them are based on Convolutional Neural Networks (CNN) (Liu et al., 2020), such as Super-Resolution CNN (SRCNN) (Dong et al., 2014). As surveyed in (Wang et al., 2020), these SISR models use several network design techniques, such as residual learning in Very Deep Super-Resolution (Kim et al., 2016), attention mechanism in Residual Channel Attention Network (RCAN) (Zhang et al., 2018), encoder-decoder network or generative adversarial network (GAN), in Super-Resolution GAN (SRGAN) (Ledig et al., 2017)and Enhanced SRGAN (ESRGAN) (Wang et al., 2018), and so on. Because of their superior performance in image Super-Resolution (Wang et al., 2018; Zhang et al., 2018; Kim et al., 2016), we will examine their downscaling per-

formance based on ensemble forecast skill metric first. In addition, we will incorporate extra inputs into Eq 1.

2.2 Downscaling techniques for climate forecasts

A large body of downscaling techniques are available in the literature, from dynamical downscaling, statistical downscaling, and recent deep learning based downscaling (Liu et al., 2020).

Dynamical downscaling, via a Regional Climate Model (RCM) forced by boundary conditions from a GCM to run finer resolution simulations, provides a nice solution to improve spatial resolution, see, e.g., (Ratnam et al., 2017; Thatcher and McGregor, 2009; Manzanas, 2020). The application of such an RCM is computationally expensive, especially for high-resolution (with grid spacing around 10 km or less) simulations over large domains and downscaling performance depends on the domain of interest. Statistical downscaling has become a normal practice to remove systematic biases, adjust the uncertainty spread and restore local daily climate variability of GCM forecasts for decision-making due to its cost-effectiveness (Maraun and Widmann, 2018). Traditional statistical downscaling techniques, especially for precipitation, have been developed with their advantages and challenges for different applications, such as Model Output Statistics (MOS), Perfect Prognosis (PP), and weather generators (Maraun and Widmann, 2018; Manzanas, 2020). MOS uses statistical approaches to enhance the forecast accuracy of a GCM by linking the distribution of GCM output to the distribution of observed local-scale climate variables. A typical example is Quantile Mapping that maps the raw forecast at a given location to the corresponding quantiles of historical observations by assuming that the distribution of model-simulated data should preserve the distribution of observed data (Michelangeli et al., 2009; Li and Jin, 2020). We denote the cumulative probability distributions (CDFs, aka quantile functions) of raw forecasts x_f and observations by F_f and F_o, respectively. The QM forecast can be formulated as $x^{(QM)} = F_o^{-1}(F_f(x_f))$ where F_o^{-1} is the inverse function of F_o. In this study, we use the empirical distribution of raw forecasts and observations over a reference period as the estimates of F_f and F_o (Maraun and Widmann, 2018). QM often has a nice performance and is often used in operation (Bureau National Operations Centre, 2019). The PP method is based on empirical relationships established between informative large-scale atmospheric variables (features) and local/regional variables of interest (predictands) such as

precipitation (Baño-Medina et al., 2020). The effective features vary from one location to another and are often required to be extracted before downscaling. Such a feature extraction procedure could capture little useful information beyond our prior knowledge, and is very time consuming as GCM's outputs are high-dimensional (Pan et al., 2019; Manzanas et al., 2018). Weather generators use stochastic models to generate Monte-Carlo time series simulations resembling the observed statistical characteristics, which explicitly model temporal dependence (Katz et al., 2003; Shao et al., 2016). These traditional downscaling methods often perform on a single grid point, and may be time-consuming, especially for ensemble SCFs with 10+ members. Comparisons of traditional statistical and dynamical downscaling suggest that neither group of methods is superior, however, in practice computationally cheaper statistical methods are widely used (Baño-Medina et al., 2020; Bureau National Operations Centre, 2019).

Pre-engineering features and relationships before the statistical downscaling modelling process limits their ability to capture important information beyond what we have known (Baño-Medina et al., 2020). Automatic feature extraction and selection integrated into the modelling process with deep learning, especially convolutional neural networks (CNN), has achieved notable success in modelling data with spatial context, recently in climate science (Reichstein et al., 2019). Deep learning has successfully been used in precipitation nowcasting (Luo et al., 2021) and precipitation parameterisations from GCMs (Pan et al., 2019). More related to our objective, downscaling techniques are developed based on SISR models (Wang et al., 2020). As one of the first works, DeepSD is proposed, which stacked SRCNN together for climate projection downscaling (Vandal et al., 2017). For long-term climate projection, (Rodrigues et al., 2018) proposed a very deep CNN-based SISR strategy to interpolate low resolution 125km weather data to 25km output for weather forecasts. (Baño-Medina et al., 2020) assessed CNN methods with three convolutional layers followed by different connection layers for downscaling 200km reanalysis precipitation to 50km observational grids over the whole of Europe. (Wang et al., 2021) proposed, based on a deep convolutional neural network with residual blocks and batch normalisation, Super-Resolution Deep Residual Network (SRDRN) for downscaling daily precipitation and temperature. SRDRN leaves behind the bias correction required in downscaling (Wang et al., 2021). YNet 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 means of precipitation (Liu et al., 2020), which has different characteristics to daily precipitation. These pioneering downscaling techniques have varying success. To reuse SISR techniques, most works oversimplified the downscaling problem by taking the upscaled observation rather than GCMs data as the input, such that the low-resolution input images match quite well with their corresponding high-resolution images, and leave behind bias-correction which is inherent in downscaling (Vandal et al., 2017; Rodrigues et al., 2018; Wang et al., 2021). Some works use relatively shallow convolution layers and their downscaling performance may not be better than classic downscaling approaches (Baño-Medina et al., 2020; Pan et al., 2019). We will select a suitable deep learning model from various SISR models based on CRPS, a widely used skill metric for ensemble forecasts. Based on the selected very deep learning structure, we incorporate other resolved climate variables and propose VDSD to enhance its downscaling performance.

3 Climate data and pre-processing

3.1 ACCESS-S1 retrospective forecast raw data

We use daily rainfall 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). It is used for climate outlooks on multi-week through to seasonal timescales. Its development is based on the United Kingdom Met Office's Global Seasonal forecast system version 5 model configuration 2 (GloSea5-GC2). ACCESS-S1 couples the state-of-the-art land surface model, ocean model and atmosphere model. Its atmosphere model has enhancements to the ensemble generation strategy to make it appropriate for sub-seasonal forecasting, and a larger ensemble size. The resolution of the atmospheric model is raised to 0.6°, (nearly 60×60km), as the Stochastic Kinetic Energy Backscatter scheme (Bowler et al., 2009) is adopted, which leads to irreparable grid-scale perturbations (MacLachlan et al., 2015). The hindcast[1] data of ACCESS-S1 are available to the public[2] from 1990 to 2012 (i.e. 23 years). In each year, it has forecasts on 48 different initialisation

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

[2] http://poama.bom.gov.au/general/hindcast_data.html.

dates (i.e. 1st, 9th, 17th, and 25th of each calendar month). Its forecasts have a lead time of 0-216 days, and 11 ensemble members. Each forecast member provides a full description of the evolution of weather for upcoming 217 days, and collectively these ensemble forecasts indicate the likelihood of a range of future weather scenarios. Daily precipitation data from ACCESS-S1 are based on the BoM's day definition of 9 am to 9 am (local time). Three precipitation forecast images for 7 Jan 2012 are illustrated in the second column, Fig 1.

ACCESS-S1 data also provides a calibrated version. As described in (Bureau National Operations Centre, 2019), for each forecast initialisation date, lead time, and grid point location, it has a calibrated function to downscale to 5km resolution. For a given forecast day, the calibration functions first carry out spatial interpolation using bilinear interpolation to high spatial resolution, and then applies QM to adjust the bias and spread between observations and forecasts in the other 22 years. We use the calibrated data for forecast skill assessment and compare them with our downscaling techniques. As its core technique is quantile mapping, we use QM to indicate it.

3.2 BARRA Reanalysis Data

The Bureau of Meteorology 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 to global reanalysis models, BARRA is expected to provide an improved understanding of the past weather than previously possible. It covers all of Australia, New Zealand and the maritime continent, and reaches a good tradeoff between the spatial resolution and consistency with precipitation observations, as recently assessed in (Acharya et al., 2019). Its spatial resolution of 0.12°, around 12km×12km, is realised in the whole region of Australia and New Zealand. BARRA use the unified model (Davies et al., 2005), a widely used grid-point atmospheric model. The model uses a complex kinetic atmospheric formula that is non-fluid and compressible, which involves the conservation of mass, time-integration method, etc. Compared with station observations, frequency distributions, extreme

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

values, and actual space-dependent and time-dependent variability can be well represented in the BARRA reanalysis data (Acharya et al., 2019). The BARRA data starts from 1 Jan 1990 to 28 Feb 2019. 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 24-h window. All of the daily aggregation is based on the same day definition of ACCESS-S1 data.

3.3 Preprocessing

We choose a region from 9°S to 43.7425°S and 112.9°E to 154.25°E as our study region, which covers the whole 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 different value ranges. For example, precipitation ranges from 0 to 900mm per day, geopotential height at 850 hPa ranges from 1200 to 1600 meters. To bring climate variables to have similar value ranges during learning, we carry out simple linear normalisation to bring each climate variable to be within [0,1].

The raw forecast data from ACCESS-S1 atmospheric grids are around 60 km. To facilitate 4-time image superresolution, we generate two versions via bicubic interpolation. One is 48km and the second is 12km. These two upsampled versions are used as inputs for SISR models or our proposed downscaling model. We pair the ACCESS-S1 forecasts made on date i with lead time l days with the BARRA reanalysis data on date $d(= i+l)$ together for training or validation. There are more than 2.5 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 for training.

4 Deep Learning for Ensemble Forecasts

We develop our deep learning technique for ensemble SCFs in two phases. First, we train several SISR techniques to generate a high-resolution image from each low-resolution precipitation forecast ensemble member, and choose the one with the best average overall forecast skill across the whole of Australia by testing it on a separate validation data set. Secondly, this most suitable SISR technique is further developed to include some auxiliary data for better forecast performance. These two steps are described in the

following two subsections respectively.

4.1 Model selection for downscaling ensemble daily forecasts

GCMs are always a substantial simplification of the real-world climate system, which is complex and high-dimensional (Maraun and Widmann, 2018). The 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 capture forecast uncertainty, ensemble forecasting becomes an operation standard for long lead time climate forecasts where multiple trajectories are provided at 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 $\hat{Y}^{(i,l,e)} \equiv \left\{\hat{y}_{j,k}^{(i,l,e)}\right\}_{m \times n}$ be precipitation raw forecast and its associated downscaled forecast, respectively, with lead time l days, ensemble number e $(e = 1, 2, ..., E)$ 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, $E = 11$, $l = 0, ..., 216$ days, $j = 1, \cdots, 316$, and $k = 1, \cdots, 376$. Fig 1 illustrates three raw precipitation forecast members from ACCESS-S1 for forecasts made on 1 Jan 2012 with lead time of 6 days. All the ensemble members target at the same date, Jan 7 Dec 2012, and share the same target images. The forecast accuracy metrics such as Mean Absolute Error (MAE), $\frac{\sum_{j,k}\left|\hat{y}_{j,k}^{(i,l,e)} - y_{j,k}^{(d)}\right|}{\sum_{j,k} 1}$, and Root Mean Square Error (RMSE), $\frac{\sum_{j,k}\left(\hat{y}_{j,k}^{(i,l,e)} - y_{j,k}^{(d)}\right)^2}{\sum_{j,k} 1}$ are not enough, especially considering possible bias and displacement in each ensemble forecast member. The Continuous Ranked Probability Score (CRPS), generalising the MAE, and is one of the most widely used overall forecast skill metric where probabilistic or ensemble forecasts are involved. It is a surrogate measure of forecast reliability, sharpness and efficiency (Grimit et al., 2006; Hersbach, 2000). It is defined as $CRPS\left(\hat{y}_{j,k}^{(i,l,e)}, y_{j,k}^{(d)}\right) = \int_{s=0}^{1}\left(F_{j,k}^{(i,l)}(s) - \mathbb{I}\left(s \leq y_{j,k}^{(d)}\right)\right)^2 ds$ where $F_{j,k}^{(i,l)}(s)$ is a (often empirical) cumulative distribution function derived from an ensemble

forecast $\left\{\hat{y}_{j,k}^{(i,l,e)}\right\}_{e=1,\cdots,E}$ and $\mathbb{I}$ is an indicator function, which represents the exceedance of the forecast compared to the actual observation $y_{j,k}^{(d)}$. $CRPS$ 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 it increases with forecast bias and spread of the ensemble forecast.

As the initialisation conditions vary from one initialisation day to another, these ensemble members do not have fixed corresponding relationship. Instead of generating an aggregated forecast from an ensemble of forecasts like 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; Jin et al., 2022). Thus, our downscaling problem can be defined as follows. For low resolution output images from GCMs, precipitation surface $X^{(i,l,e)} \in \mathcal{R}^{m_0 \times n_0}$ and other climate variable surfaces $Z^{(i,l,e)} \in \mathcal{R}^{m_0 \times n_0 \times p}$ with respect to 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 $Y^{(d)}$, $\hat{Y}^{(i,l,e)} = \left\{\hat{y}_{j,k}^{(i,l,e)}\right\}_{m \times n} = \mathcal{G}\left(X^{(i,l,e)}, Z^{(i,l,e)}; \theta\right)$, that can minimise the average CRPS across all the validation image pairs:

$$\overline{CRPS} = \frac{\sum\limits_{i,l,j,k} w_{j,k}^{(i,l)} CRPS\left(\hat{F}_{j,k}^{(i,l)}, y_{j,k}^{(d)}\right)}{\sum\limits_{i,l,j,k} w_{j,k}^{(i,l)}} \tag{2}$$

where $\hat{F}_{j,k}^{(i,l)}$ is the empirical cumulative distribution function estimated from $\left\{\hat{y}_{j,k}^{(i,l,e)}\right\}$ for $e = 1, \cdots, E$, and $w_{j,k}^{(i,l)}$ is the weight for the ensemble forecast made on date i, lead time l at location (j,k), and $d = i + l$. We use $w_{j,k}^{(i,l)}$=1 for this study for simplicity.

To locate such a good function $\mathcal{G}$ and its parameter θ, 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 insert extra variable $Z^{(i,l,e)}$ to enhance its downscaling performance. We partition all the initialisation dates in the 23 years randomly into two groups. The first group has 1056 initialisation dates and image pairs from this group is used for model training. The image pairs from the remaining 48 initialisation dates

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

$$\hat{\theta} = \arg\min_{\theta} \left[\mathcal{L}\left(\hat{Y}^{(i,l,e)}, Y^{(d)} \right) + \lambda \Phi(\theta) \right], \tag{3}$$

L is the loss function, calculating an error between high-resolution images and super-resolution image (output from of Eq 1); λ is tradeoff parameter and $\Phi(\theta)$ is the regularisation term.

On the separate validation data sets, the average CRPS skill score of trained ESRGAN across the whole Australia is lower than QM, especially for leading time up to 30 days. For our downscaling data set, RCAN is found relatively hard to converge. It is partially because its cross-channel dependency mechanism became useless for our data. Another possible reason is that bias and displacement are prevalent in our climate image pairs. The validation forecast skill of trained RCAN is not as good as QM either. However, VDSR is much faster to converge and outperforms QM in terms of average CRPS. We select the VDSR model for further development. We also try a few different settings of VDSR, such as with 8, 12, 15, 18, 30, or 36 layers of convolution and activation. Its validation CRPS decreases from 8 to 18 layers, and after that, did not change that much. We stick with 18 layers as recommended by (Kim et al., 2016) for SISR.

4.2 Very Deep Statistical Downscaling (VDSD)

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 (VDSD). The climate variables, different from precipitation, are resolvable in the climate modelling and often have more reliable forecasts (Pan et al., 2019; Baño-Medina et al., 2020; Merryfield et al., 2020). The overall structure of finalised VDSD is illustrated in Fig **??**, where Geopotential Height (GZ) at

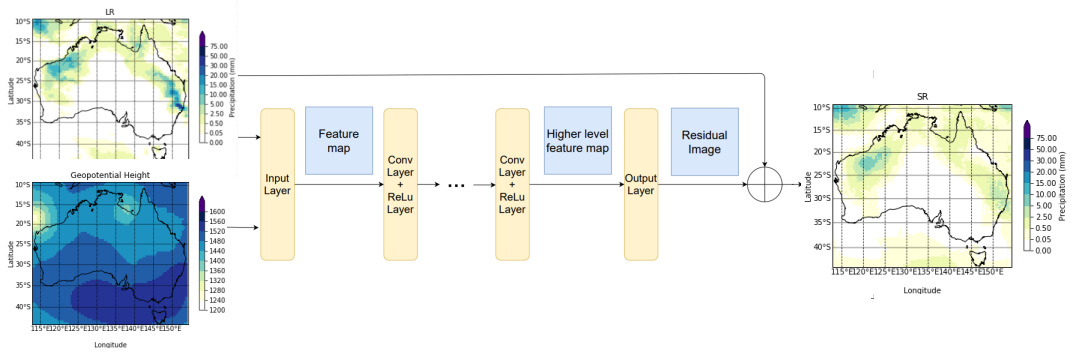


Fig. 2: The structure of the VDSD model, modified from VDSR (Kim et al., 2016), where $\oplus$ represents element-wise matrix addition with input precipitation image, orange blocks are layers of the neural network, blue rectangles are feature maps, input and output climate data images are on the left-hand and right-hand side respectively. These input/output images are shown in the original scale, instead of the normalised scale between [0,1].

850 hPa is used as the additional input. This GZ represents the altitude above mean sea level at which the atmospheric pressure is 850 hPa. It is influenced by both temperature and air pressure, and is a reliable output from climate models.

VDSD, modified from VDSR, mainly has three parts: input, intermediate feature extraction, and output layers. It can take precipitation images and other climate images as input. These input images have been pre-processed with the same spatial resolution as high-resolution output images, and the same value range between 0 and 1 (detailed in Section 3.3). These input images go through multiple feature extraction layers. 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, it finally generates an output image at the same resolution as the label image. VDSD maintains the residual learning which has been widely demonstrated on robust and speedy training in SISR (Kim et al., 2016; Wang et al., 2020).

As shown in Fig **??**, two or more input images X_{lr} and Z_{lr}, which represent the raw climate forecasts after upsampling, firstly go through the input layer. This layer has a convolution layer and a ReLU layer. The convolution layer

has 64 kernels and produces 64 first-level feature maps. Then the ReLU layer performs the ReLU function to force negative values from the feature maps to be zero. The operation can be formulated as

$$M_0 = B\left(X_{lr}, Z_{lr}\right) = ReLU\left(Conv\left(X_{lr}, Z_{lr}\right)\right) \tag{4}$$

where M_0 is the first level feature maps generated by input layer, and $ReLU()$ and $Conv()$ are ReLU and convolution layer that perform the ReLU function and 2-dimensional convolution. Each convolutional layer has 64 kernels to produce 64 feature maps. The kernel size is set to be 3x3. Padding and the step length are 1. Therefore, the size of each feature map is the same as the size of high-resolution image. Suppose the size of input images is $2 \times m \times n$, then the size of the feature map generated by the input layer is $64 \times m \times n$. 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 features, and a ReLU layer, which forces negative values to be zero. Each convolutional layer takes 64 feature maps from the previous block as input. Therefore, the operation of each intermediate block is the same, which can be written as

$$M_t = B\left(M_{t-1}\right) = ReLU\left(Conv\left(M_{t-1}\right)\right) = B^t\left(M_0\right) \tag{5}$$

where M_t represents the tth level feature map, B is the operation of an intermediate block. These intermediate blocks can extract higher-level features from extra climate variables too to capture complex patterns, which are expected to improve downscaling performance.

The output layer is a convolutional layer that converts 64 high level feature maps into a residual image – that is to use discovered complex patterns 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.

We again use the average CRPS across the whole Australia on the separate validation data set to finalise structure of VDSD. We test two different types of variants of VDSD. (1) One is adding extra input images for downscaling. We try four climate variables from ACCESS-S1 and their combinations. They are geopotential height (at 850 hPa), daily maximum temperature, daily minimum temperature, and sea level pressure. Our results shew that adding more climate variables than geopotential height rarely improved CRPS, and sometimes deteriorated its ensemble forecast skills. (2)

The other is to try two different loss functions in 3, i.e., L1, $\mathcal{L}\left(\hat{Y}^{(i,l,e)}, Y^{(d)}\right) = \frac{\sum_{k,j}\left|\hat{y}_{k,j}^{(i,l,e)} - y_{k,j}^{(d)}\right|}{\sum_{k,j} 1}$, or L2, $\mathcal{L}\left(\hat{Y}^{(i,l,e)}, Y^{(d)}\right) = \frac{\sqrt{\sum_{k,j}\left(\hat{y}_{k,j}^{(i,l,e)} - y_{k,j}^{(d)}\right)^2}}{\sum_{k,j} 1}$. L1 gives us better validation CRPS, and would be used in our final VDSD model. Note that in VDSR, L2 is preferred.

5 Cross-validation results and comparison

To illustrate the downscaling performance of VDSR and VDSD that we've finalised, we used the last few years hindcast data for cross-validation. We conducted two leave-one-year-out validations. (1) We took forecasts made on 48 initialisation dates in 2012 for validation and the other forecasts made before 2012 for training the downscaling methods. Daily BARRA precipitation data between 1 Jan 2012 and 29 July 2013 were used in validation as the ACCESS-S1 forecasts made on 25 Dec 2012 cover up to 29 July 2013 for its 216-days lead time forecasts. (2) We left forecasts made on 48 initialisation dates in 2010 as validation 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 cross-validation. Thus around 1149 precipitation images were used in cross-validation.

5.1 Performance Metrics for Forecasts

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

As we discuss in Section 4.1, the average CRPS of ensemble forecasts for each grid point on validation data is treated as an overall ensemble forecast skill assessment. For each grid point, averaging across all the initialisation dates in the validation period, we obtain the averaged CRPS of a forecast model for a lead time. For further comparison with climatology and easy understanding, we calculate the CRPS skill score for model m against the CRPS of climatology as follows.

$$CRPS\_SS^{(m)} = 1 - \frac{\overline{CRPS}^{(m)}}{\overline{CRPS}^{(clim)}} \tag{6}$$

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 with its associated observation. The skill score is zero if a forecast has identical CRPS as climatology. A positive CRPS skill score indicates the downscaled forecast is better than the climatology model, and vice versa.

As downscaling techniques based on deep learning 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 $\overline{MAE}_{l,k,j} = \frac{\sum_{f,e}\left|\hat{Y}_{k,j}^{(f,l,e)} - Y_{k,j}^{(d)}\right|}{\sum_{f,e} 1}$ for lead time of l days. Taking climatology as the reference forecast, we can define MAE skill score for model m for each pixel as

$$MAE\_SS_{l,k,j}^{(m)} = 1 - \frac{\overline{MAE}_{l,k,j}^{(m)}}{\overline{MAE}_{l,k,j}^{(clim)}} \tag{7}$$

Similarly, a higher MAE skill score is preferred. For a fair comparison, skill scores presented in the following exclude locations on the ocean as the QM model focused on Australian land.

5.2 Results for forecasts made on 48 initialisation days in 2012

As illustrated in Fig 1 for three ensemble members forecasted on 1 Jan 2012 for 7 Jan 2012, VDSR keeps similar precipitation area patterns as spatial-interpolated ACCESS-S1 raw forecasts and often has more areas with precipitation. Downscaled images of VDSD follow precipitation shapes of the raw forecasts while more likely reduces precipitation amount for this day, which may cause issues for wet days. VDSD could adjust the precipitation area shapes. For ensemble members 2 and 11, VDSD substantially reduces the precipitation along the 30°S latitude line, which brings its downscaled images closer to the observations for 7 Jan 2012.

Averaging across 48 initialisation days in 2012, we calculate its average CRPS value for each grid point. Fig 3 illustrates mean CRPS skill scores

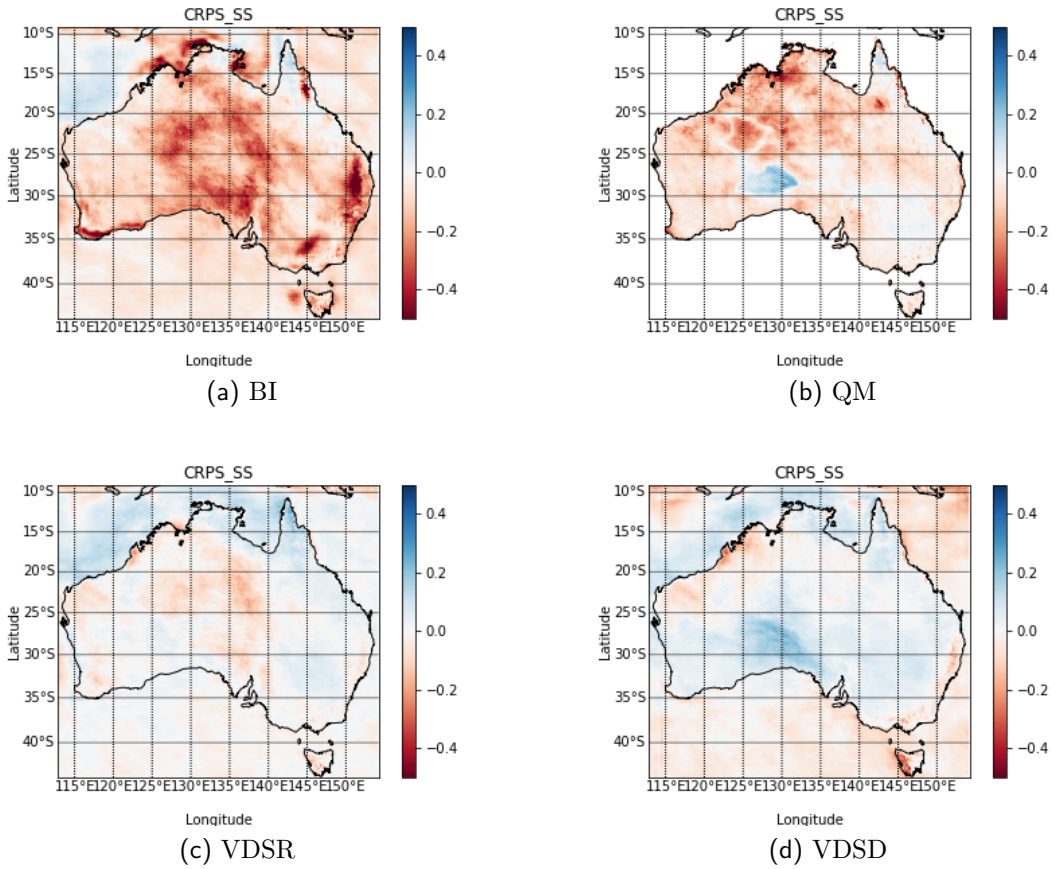
across the whole Australia of four downscaling models along with lead time up to 216 days. VDSR has the highest scores in the first three lead times, and then VDSD becomes the best of four models for most other lead times. For example, for the lead time of 6 days (some typical downscaling results are illustrated in Fig 1), its CRPS skill scores for the four models are spatially visualised in Fig. 9 [4]. For most locations on Australian land (except north-western Australia, eastern seaboard of Australia, and Tasmania), VDSD 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 VDSD 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 (Fig 3). Averaging across the 217 different lead times, VDSR has positive CRPS skill scores for most locations on Australian land (except north-western Australia, eastern seaboard of Australia, and Tasmania), while its three counterparts have negative skills for most locations (Fig. 10). That means only VDSD is better than the climatology for most locations on Australian land. Their mean CRPS skill scores are 5.63×10^{-3}, -2.54×10^{-2}, -1.05×10^{-1}, and -1.42×10^{-1} respectively, for VDSD, VDSR, QM and BI. Among the four downscaling techniques, only VDSD is better than climatology on average as its positive CRPS skill score. VDSD is about 0.03 better than VDSR, and more than 0.11 better than both traditional downscaling techniques. Along with 217 different lead times, the correlation of VDSD's skill scores with these of VDSR and QM is around 0.71, and 0.33 with BI. The correlation between VDSR and QM is 0.86. The lower correlation between VDSD with QM than VSDR with QM is likely caused by including the extra climate variable geopotential height.

To check the performance of these downscaling techniques for sub-seasonal forecasts, Fig 4 illustrates the average CRPS skill scores for the first 45 lead times. The skill scores of BI are around their mean -1.39×10^{-1} for most locations on Australian land. QM has some improvement to be 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}. VDSD has positive skill scores for most locations on Australian land, and its average is around 2.76×10^{-2}. VDSD still has negative skill scores along the eastern coastline and north-western parts of Australia, as well as Tasmania.

Fig 5 illustrates the MAE skill scores of four downscaling techniques along

[4] To facilitate an easy comparison, these spatial plots use the same colour bar.

Fig. 3: Average CRPS Skill Scores across Australia for the forecasts made in 2012.



(a) BI (b) QM

(c) VDSR (d) VDSD

Fig. 4: Average CRPS skill score for lead time 0 to 44 days across Australia for forecasts made in 2012

lead times. Except for the first six lead times, BI has negative skill scores, which indicates the ACCESS-S1 raw forecasts have limited skill. QM often has positive MAE skill scores. Both deep learning models, VDSR and VDSD have substantial improvement for all the different lead times. Averaging across these 217 different lead times, the MAE skill scores of the four models are around 0.38, 0.19, 0.02, and -0.13 respectively.

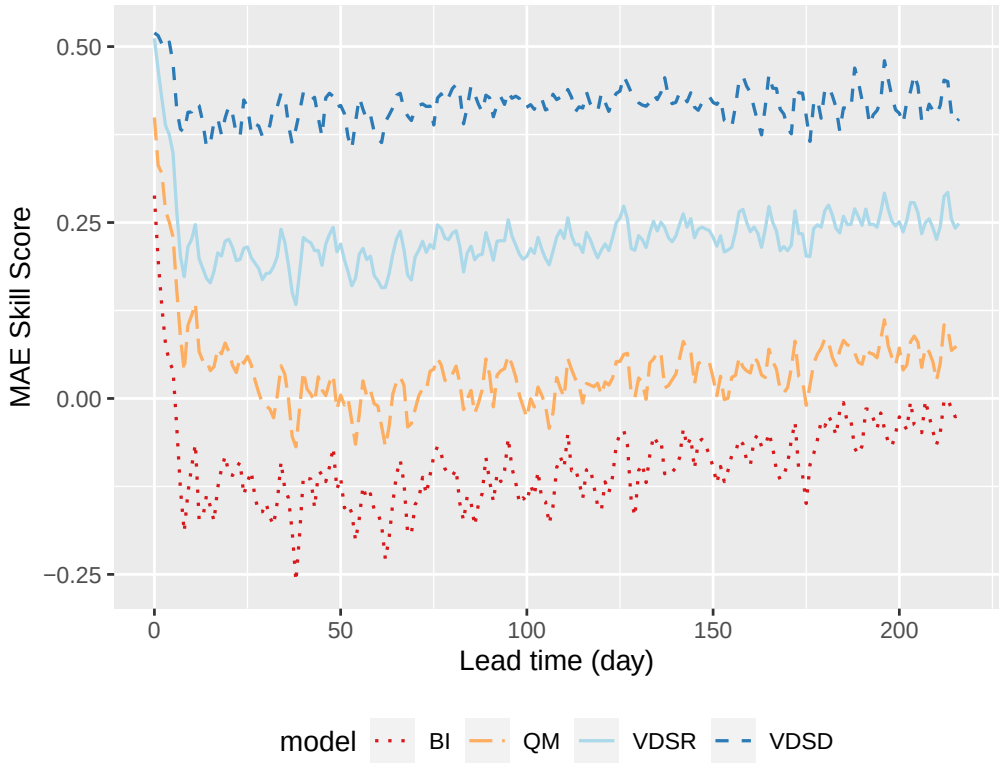


Fig. 5: Average MAE skill scores across Australia for daily precipitation forecasts made on 48 different initialisation dates in 2012

5.3 Results for forecasts made on 48 initialisation days in 2010

Fig 6 illustrates mean CRPS skill scores along with lead time based on 48 seasonal ensemble forecasts made in 2010. For most time of the 217 different lead times, VDSD 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 on most locations on Australian land. In comparison, QM and BI have negative CRPS skill scores on a lot of locations (Fig. 11). VDSD has higher skill scores than VDSR in southeast and south central Australia though both have quite similar spatial patterns. On average, the average CRPS skill scores of VDSD and VDSR are 5.34×10^{-2} and 3.97×10^{-2}. They are much higher than -3.01×10^{-2} and -8.79×10^{-2} of QM and BI respectively (Fig. 6). The average CRPS skill scores across the 217 different lead times of VDSD and VDSR are often positive or close to zero for most locations on Australian land (Figs 12d and 12c), and both QM and BI are normally in the negative domain (Figs 12a and 12b). The mean CRPS skill scores for VDSD, VDSR, QM 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. 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, instead of 22 in climatology, ensemble members which could decrease about a few per cent of CRPS skill score (Ferro et al., 2008; Li and Jin, 2020). For the first 45 different lead times, the mean CRPS skill scores are 1.38×10^{-2}, -1.02×10^{-3}, -6.62×10^{-2}, -9.06×10^{-2}, respectively, for VDSD, VDSR, QM and BI. As illustrated in Fig 7, for most locations on Australian land, both VDSR and VDSD have between -0.1 and 0.1 CRPS skill score though VDSD has slightly higher CRPS skill scores in northern and eastern Australia.

Fig 8 illustrates the MAE skill scores of four downscaling techniques along lead times. Except for the first eight lead times, both BI and QM have negative skill scores. VDSR and VDSD always have positive skill scores. Averaging across these 217 lead times, the MAE skill scores of these four models are -0.19, 0.09, 0.21 and 0.24 respectively. VDSD has a relatively small improvement against VDSR, and both are much better than climatology. Considering these, we think VDSD is comparable with climatology in terms of both forecast accuracy and ensemble forecast skill for SCFs made

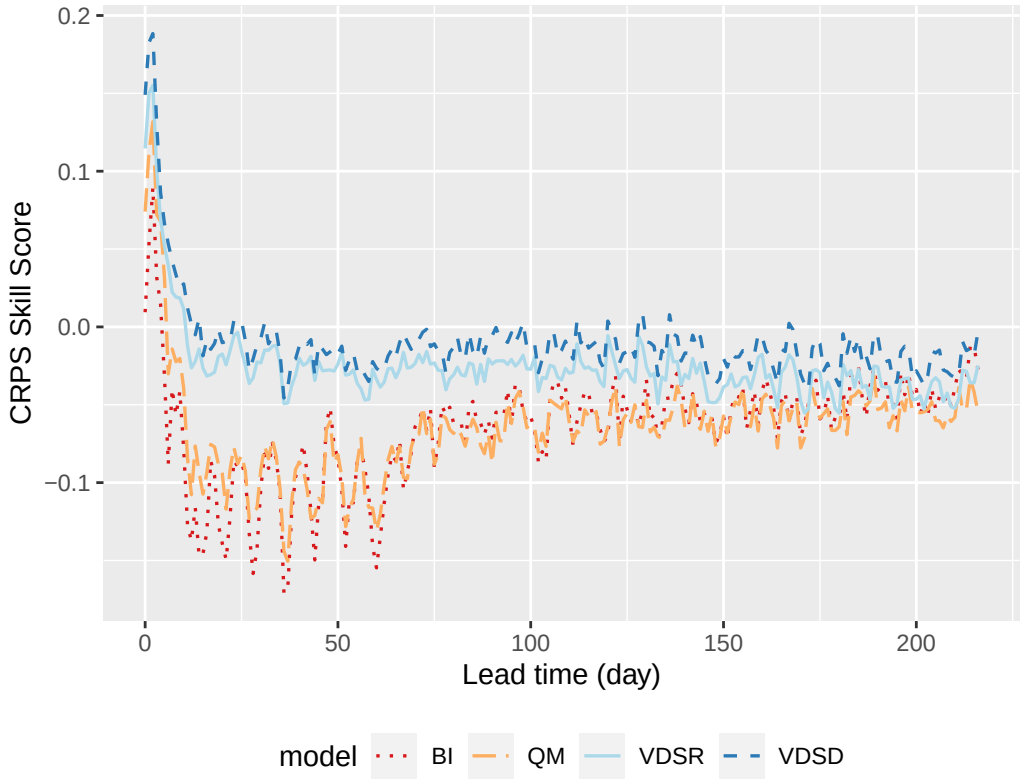
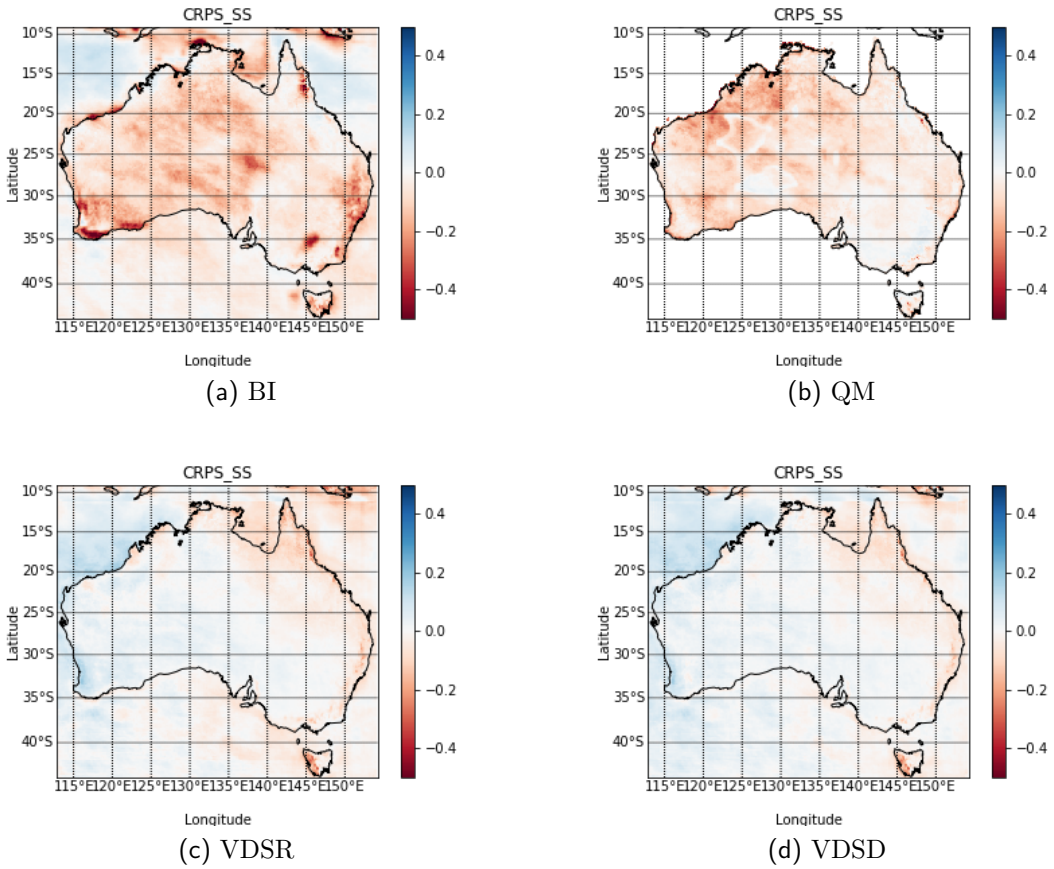


Fig. 6: Mean CRPS skill scores across Australian land for forecasts made in 2010.



(a) BI (b) QM

(c) VDSR (d) VDSD

Fig. 7: Average CRPS skill score for lead time 0 to 44 days across Australia for forecasts made in 2010

on the 48 initialisation dates in 2010.

5.4 Implementation and computation time

The hyper-parameters used in the training of both VDSR and VDSD are as follows. The number of epochs was 50, for which we saw the objective function stabilised. The learning rate was 0.0001, and relatively small as the network is very deep, and a large learning rate may cause vanishing/exploding gradients problem (Bengio et al., 1994). The optimisation method is stochastic gradient descent with momentum of 0.9. Our implementation is written

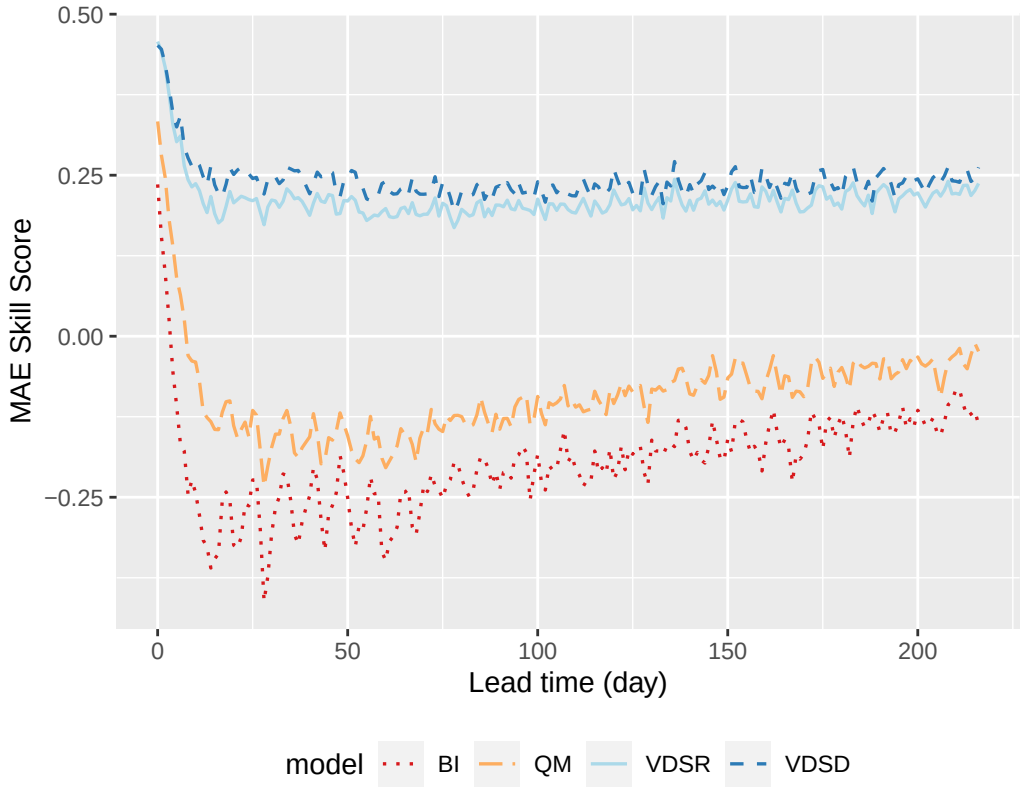


Fig. 8: Mean MAE skill scores for whole Australia for 2010-2011 precipitation forecast.

	Gadi in NCI.org.au	PC
CPU	36 × Intel® Xeon^{TM} Platinum 8268	1× Intel® Core^{TM} i5-9600K
CPU clock rate	2.9 GHz	3.70 GHZ
CPU logical cores	36	6
CPU cache	35.75 MB	9 MB
GPU	3× Nvidia® V100	GeForce RTX 2070
GPU memory	32 GB	8 GB
CUDA(R) cores	5120	2304

Tab. 1: Hardware configuration used in the experiments

in python (v3.7.4).

Table 1 lists the hardware we used in the experiments. The training was run on Gadi (the second column), a high-performance computer in National Computational Infrastructure (NCI), Australia. Forecast validation was done on a normal PC with a GPU.

Table 2 lists average computation time required for both training and operation where 11 ensemble members for 217 days rainfall forecasts from ACCESS-S1 were downscaled. The total training time for optimising VDSD model parameters was around 16.76 hours, which is about 38% longer than VDSR. BI and QM don't require training time. Downscaling operation for a single seasonal forecast was run on a PC, BI, QM, VDSR and VDSD required 0.02, 11.21, 0.08 and 0.56 hours. VDSD is 7 times slower than VDSR and 20 times faster than QM.

Another statistical downscaling technique Extended Copula-based Post-Processing (ECPP)(Li and Jin, 2020) needed about one hour for training for each station or grid point and took 0.46 seconds for an operation forecast. It would take around 15.1 hours to downscaled rainfall for the whole Australian land on a normal PC. The dynamic downscaling model, Conformal Cubic Atmospheric Model (CCAM)(Thatcher and McGregor, 2009), don't need training time, and took about 0.33 hours to simulate a single 1-month lead time forecast to 10km resolution on a CSIRO supercomputer Pearcey with 1536 cores (personal communication with M. Thatcher). Compared with ECPP and CCAM, VDSD is much faster for seasonal rainfall downscaling operation.

Tab. 2: Computation time in hours of four downscaling methods

Method	Training time on Gadi	Operation time on a normal PC
BI	0	0.02
QM	0	11.21
VDSR	12.12	0.08
VDSD	16.76	0.56

6 Conclusion and Future Work

To improve the downscaling techniques for long lead time daily probabilistic precipitation forecasts in Australia, we have applied several representatives Single Image Super-Resolution (SISR) techniques and selected Very Deep Super-Resolution (VDSR) as the suitable deep learning model. The selection has been based on the overall probabilistic forecast skill metric – Continuous Ranked Probability Score (CRPS) on a separated validation data set. We have further incorporated extra climate variables into VDSR and established the Very Deep Statistical Downscaling (VDSD) model. Both deep learning models have been finalised their structures based on CRPS on the validation data set. 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 in terms of both forecast accuracy and CRPS. VDSD have outperformed climatology, a benchmark for long lead time ensemble climate forecast, in 2012 and the first 15 lead times in 2010. Both VDSR and VDSD have downscaled long lead time daily precipitation very fast while demanding a lot of time for model development and training. Thus, deep learning models, especially, the proposed VDSD, have demonstrated their potential for possible operational use in future.

For validation results for forecasts made in 2010, the overall average ensemble forecast skill of VDSD is slightly worse than climatology. There are three possible reasons. (1) For forecasts made in 2010, rainfall data from 1 Jan 2010 to 29 July 2011 are used for skill assessment. Years 2010 and 2011 are the third-wettest 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 [5]. The La Niña event peak in 2012 is much weaker, and made 2012 relatively easier

[5] (http://www.bom.gov.au/climate/enso/lnlist/)

to forecast. That means the training data for models tested on 2011 have relatively less precipitation, hence VDSD intends to move in that direction, which deteriorated its performance for forecasts made in 2010. (2) The host climate model ACCESS-S1 may perform worse in 2010 than 2012, on e.g., geopotential heights. For both validation settings, VDSD has substantial improvement from the raw forecasts from the climate model ACCESS-S1, and its final performance still heavily depends on ACCESS-S1's raw forecasts. (3) The climatology benchmark we have used has 22 ensemble members, and a double ensemble size led to a few per cent higher on CRPS (Ferro et al., 2008). Therefore, although the CRPS skill score is negative on average, the proposed VDSD is thought to be comparable with climatology. It is still promising for operation because it needs less downscaling operation time.

Though deep learning models can provide more skilful high-resolution continuous SCFs to drive impact models or biophysical models, the accuracy and skills of these SCFs may not be high enough for direct use in wider communities such as agriculture and hydrology (Kusunose and Mahmood, 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 be not very consistent with on-the-ground observations (Acharya et al., 2019). To remove such inconsistency, station-specific downscaling techniques like QM, ECPP and their variants (Li and Jin, 2020) may further improve long lead time forecasts. As the spatial and cross-variable relationships may be not stationary, we will investigate separate downscaling models for different seasons, which is often very helpful in practice. Our cross-validation assessment has not considered extreme rainfall events closely, which have a huge impact on real applications (Li et al., 2021). VDSD only downscales to 12km, which should be further increased for applications. For a fair comparison and reducing training time, we have only used the forecasts with lead times less than seven days as training data. That may lead to putting more emphasis on low-resolution precipitation and less on correcting inherent bias of GCM's outputs. A tradeoff between bias correction and resolution improvement is also subject to future work.

Acknowledgements

This work was partially funded by the CSIRO Digiscape Future Science Platform.

References

W. J. Merryfield, J. Baehr, L. Batté, E. J. Becker, A. H. Butler, C. A. Coelho, G. Danabasoglu, P. A. Dirmeyer, F. J. Doblas-Reyes, D. I. Domeisen *et al.*, "Current and emerging developments in subseasonal to decadal prediction," *Bulletin of the American Meteorological Society*, vol. 101, no. 6, pp. E869–E896, 2020.

R. Manzanas, "Assessment of model drifts in seasonal forecasting: Sensitivity to ensemble size and implications for bias correction," *Journal of Advances in Modeling Earth Systems*, vol. 12, no. 3, p. e2019MS001751, 2020.

K. A. Parton, J. Crean, and P. Hayman, "The value of seasonal climate forecasts for Australian agriculture," *Agricultural Systems*, vol. 174, pp. 1–10, 2019.

J. W. Mjelde and J. F. Griffiths, "A review of current evidence on climate forecasts and their economic effects in agriculture," *American Journal of Agricultural Economics*, vol. 80, no. 5, pp. 1089–1095, 1998.

The Centre for International Economics, "Analysis of the benefits of improved seasonal climate forecasting for agriculture," Managing Climate Variability Program, Tech. Rep., 2014, accessed in Nov 2020. [Online]. Available: http://www.climatekelpie.com.au/Files/MCV-CIE-report-Value-of-improved-forecasts-non-agriculture-2014.pdf

Y. Kusunose and R. Mahmood, "Imperfect forecasts and decision making in agriculture," *Agricultural Systems*, vol. 146, pp. 103–110, 2016.

M. Li and H. Jin, "Development of a postprocessing system of daily rainfall forecasts for seasonal crop prediction in australia," *Theoretical and Applied Climatology*, vol. 141, pp. 1331–1349, 2020.

A. Schepen, Y. Everingham, and Q. J. Wang, "An improved workflow for calibration and downscaling of GCM climate forecasts for agricultural applications – a case study on prediction of sugarcane yield in australia," *Agricultural and Forest Meteorology*, vol. 291, p. 107991, 2020.

D. Hudson, O. Alves, H. H. Hendon, E.-P. Lim, G. Liu, J.-J. Luo, C. MacLachlan, A. G. Marshall, L. Shi, G. Wang *et al.*, "ACCESS-S1

the new bureau of meteorology multi-week to seasonal prediction system," *Journal of Southern Hemisphere Earth Systems Science*, vol. 67, no. 3, pp. 132–159, 2017.

S. Saha, S. Moorthi, X. Wu, J. Wang, S. Nadiga, P. Tripp, D. Behringer, Y.-T. Hou, H.-y. Chuang, M. Iredell *et al.*, "The NCEP climate forecast system version 2," *Journal of climate*, vol. 27, no. 6, pp. 2185–2208, 2014.

S. J. Johnson, T. N. Stockdale, L. Ferranti, M. A. Balmaseda, F. Molteni, L. Magnusson, S. Tietsche, D. Decremer, A. Weisheimer, G. Balsamo, S. P. E. Keeley, K. Mogensen, H. Zuo, and B. M. Monge-Sanz, "SEAS5: the new ECMWF seasonal forecast system," *Geoscientific Model Development*, vol. 12, no. 3, pp. 1087–1117, 2019.

T. Vandal, E. Kodra, S. Ganguly, A. Michaelis, R. Nemani, and A. R. Ganguly, "DeepSD: Generating high resolution climate change projections through single image super-resolution," in *KDD'17*, 2017, pp. 1663–1672.

J. Ratnam, T. Doi, and S. K. Behera, "Dynamical downscaling of SINTEX-F2v CGCM seasonal retrospective austral summer forecasts over australia," *Journal of Climate*, vol. 30, no. 9, pp. 3219–3235, 2017.

Y. Liu, A. R. Ganguly, and J. Dy, "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 & Data Mining*, 2020, pp. 3145–3153.

J. Baño-Medina, R. Manzanas, and J. M. Gutiérrez, "Configuration and intercomparison of deep learning neural models for statistical downscaling," *Geoscientific Model Development*, vol. 13, no. 4, pp. 2109–2124, 2020.

D. Maraun and M. Widmann, *Statistical downscaling and bias correction for climate research*. Cambridge University Press, 2018.

M. Bettolli, S. Solman, R. Da Rocha, M. Llopart, J. Gutierrez, J. Fernández, M. Olmo, A. Lavin-Gullon, S. Chou, D. C. Rodrigues *et al.*, "The cordex flagship pilot study in southeastern south america: a comparative study of statistical and dynamical downscaling models in simulating daily extreme precipitation events," *Climate Dynamics*, vol. 56, no. 5, pp. 1589–1608, 2021.

M. Thatcher and J. L. McGregor, "Using a scale-selective filter for dynamical downscaling with the conformal cubic atmospheric model," *Monthly Weather Review*, vol. 137, no. 6, pp. 1742–1752, 2009.

R. Manzanas, A. Lucero, A. Weisheimer, and J. M. Gutierrez, "Can bias correction and statistical downscaling methods improve the skill of seasonal precipitation forecasts?" *Climate Dynamics*, vol. 50, no. 3-4, pp. 1161–1176, 2018.

M. Reichstein, G. Camps-Valls, B. Stevens, M. Jung, J. Denzler, N. Carvalhais *et al.*, "Deep learning and process understanding for data-driven earth system science," *Nature*, vol. 566, no. 7743, pp. 195–204, 2019.

X. Shi, Z. Gao, L. Lausen, H. Wang, D.-Y. Yeung, W.-k. Wong, and W.-c. Woo, "Deep learning for precipitation nowcasting: A benchmark and a new model," in *Advances in neural information processing systems*, 2017, Conference Proceedings, pp. 5617–5627.

C. Luo, X. Li, and Y. Ye, "PFST-LSTM: a spatiotemporal lstm model with pseudo-flow prediction for precipitation nowcasting," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 14, pp. 843–857, 2021.

S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-c. Woo, "Convolutional lstm network: A machine learning approach for precipitation nowcasting," in *Advances in neural information processing systems*, 2015, pp. 802–810.

B. Pan, K. Hsu, A. AghaKouchak, and S. Sorooshian, "Improving precipitation estimation using convolutional neural network," *Water Resources Research*, vol. 55, no. 3, pp. 2301–2321, 2019.

E. R. Rodrigues, I. Oliveira, R. Cunha, and M. Netto, "DeepDownscale: a deep learning strategy for high-resolution weather forecast," in *2018 IEEE 14th International Conference on e-Science (e-Science)*, 2018, pp. 415–422.

F. Wang, D. Tian, L. Lowe, L. Kalin, and J. Lehrter, "Deep learning for daily precipitation and temperature downscaling," *Water Resources Research*, p. e2020WR029308, 2021.

Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 10, pp. 3365–3387, 2020.

E. P. Grimit, T. Gneiting, V. J. Berrocal, and N. A. Johnson, "The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification," *Quarterly Journal of the Royal Meteorological Society*, vol. 132, no. 621C, pp. 2925–2942, 2006.

C. A. Ferro, D. S. Richardson, and A. P. Weigel, "On the effect of ensemble size on the discrete and continuous ranked probability scores," *Meteorological Applications*, vol. 15, no. 1, pp. 19–24, 2008.

J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *CVPR*, 2016, pp. 1646–1654.

C. Dong, C. C. Loy, K. He, and X. Tang, "Learning a deep convolutional network for image super-resolution," in *European conference on computer vision*. Springer, 2014, pp. 184–199.

Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, "Image super-resolution using very deep residual channel attention networks," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 286–301.

C. Ledig, L. Theis, F. HuszÃ¡r, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, and Z. Wang, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, Conference Proceedings, pp. 4681–4690.

X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. Change Loy, "Esrgan: Enhanced super-resolution generative adversarial networks," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, Conference Proceedings, pp. 63–79.

P. A. Michelangeli, M. Vrac, and H. Loukos, "Probabilistic downscaling approaches: Application to wind cumulative distribution functions," *Geophysical Research Letters*, vol. 36, 2009.

Bureau National Operations Centre, "Operational implementation of ACCESS-S1 forecast post processing," Bureau of Meteorology, Tech. Rep. 124, Sep 2019.

R. W. Katz, M. B. Parlange, and C. Tebaldi, "Stochastic modeling of the effects of large-scale circulation on daily weather in the southeastern us," *Climatic Change*, vol. 60, no. 1-2, pp. 189–216, 2003.

Q. Shao, L. Zhang, and Q. Wang, "A hybrid stochastic-weather-generation method for temporal disaggregation of precipitation with consideration of seasonality and within-month variations," *Stochastic Environmental Research and Risk Assessment*, vol. 30, no. 6, pp. 1705–1724, 2016.

N. E. Bowler, A. Arribas, S. E. Beare, K. R. Mylne, and G. J. Shutts, "The local etkf and skeb: Upgrades to the mogreps short-range ensemble prediction system," *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, vol. 135, no. 640, pp. 767–776, 2009.

C. MacLachlan, A. Arribas, K. Peterson, A. Maidens, D. Fereday, A. Scaife, M. Gordon, M. Vellinga, A. Williams, R. Comer *et al.*, "Global seasonal forecast system version 5 (glosea5): a high-resolution seasonal forecast system," *Quarterly Journal of the Royal Meteorological Society*, vol. 141, no. 689, pp. 1072–1084, 2015.

C.-H. Su, N. Eizenberg, P. Steinle, D. Jakob, P. Fox-Hughes, C. J. White, S. Rennie, C. Franklin, I. Dharssi, and H. Zhu, "BARRA v1.0: the bureau of meteorology atmospheric high-resolution regional reanalysis for australia," *Geoscientific Model Development*, vol. 12, no. 5, pp. 2049–2068, 2019.

S. C. Acharya, R. Nathan, Q. J. Wang, C.-H. Su, and N. Eizenberg, "An evaluation of daily precipitation from a regional atmospheric reanalysis over australia," *Hydrology and Earth System Sciences*, vol. 23, no. 8, pp. 3387–3403, 2019. [Online]. Available: https://doi.org/10.5194/hess-23-3387-2019

T. Davies, M. J. Cullen, A. J. Malcolm, M. Mawson, A. Staniforth, A. White, and N. Wood, "A new dynamical core for the met office's global and regional modelling of the atmosphere," *Quarterly Journal of the Royal Me-*

teorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography, vol. 131, no. 608, pp. 1759–1782, 2005.

H. Hersbach, "Decomposition of the continuous ranked probability score for ensemble prediction systems," *Weather and Forecasting*, vol. 15, no. 5, pp. 559–570, 2000.

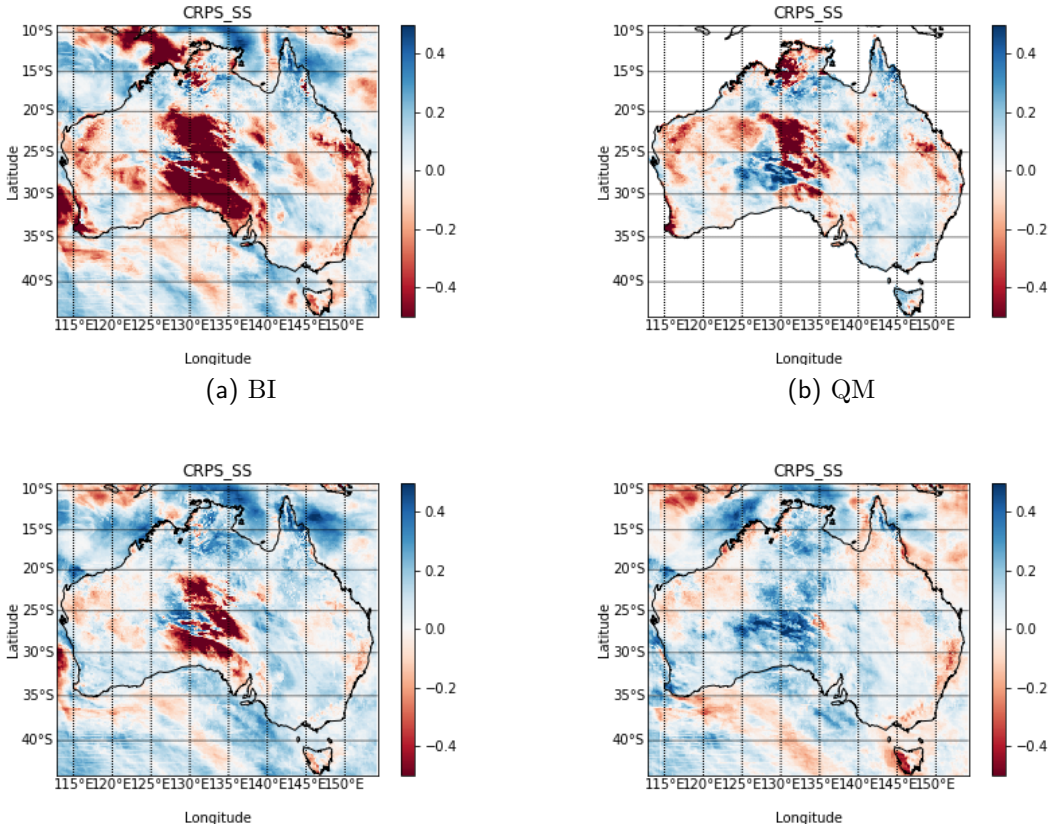
B. Basso and L. Liu, *Seasonal crop yield forecast: Methods, applications, and accuracies*. Elsevier, 2019, vol. 154, pp. 201–255.

H. Jin, M. Li, G. Hopwood, Z. Hochman, and K. S. Bakar, "Improving early-season wheat yield forecasts driven by probabilistic seasonal climate forecasts," *Agricultural and Forest Meteorology*, vol. 315, p. 108832, 2022.

Y. Bengio, P. Simard, and P. Frasconi, "Learning long-term dependencies with gradient descent is difficult," *IEEE transactions on neural networks*, vol. 5, no. 2, pp. 157–166, 1994.

M. Li, H. Jin, and Q. Shao, "Improvements in subseasonal forecasts of rainfall extremes by statistical postprocessing methods," *Weather and Climate Extremes*, vol. 34, p. 100384, 2021.

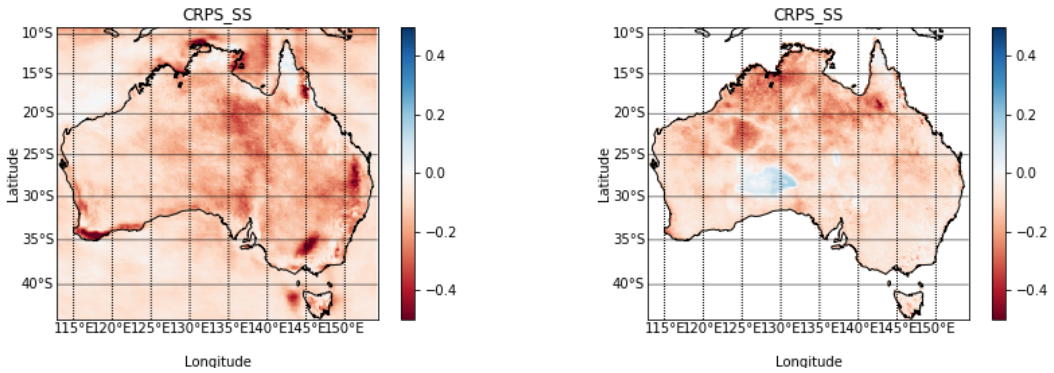
A Supplement Figures



(a) BI (b) QM

(c) VDSR (d) VDSD

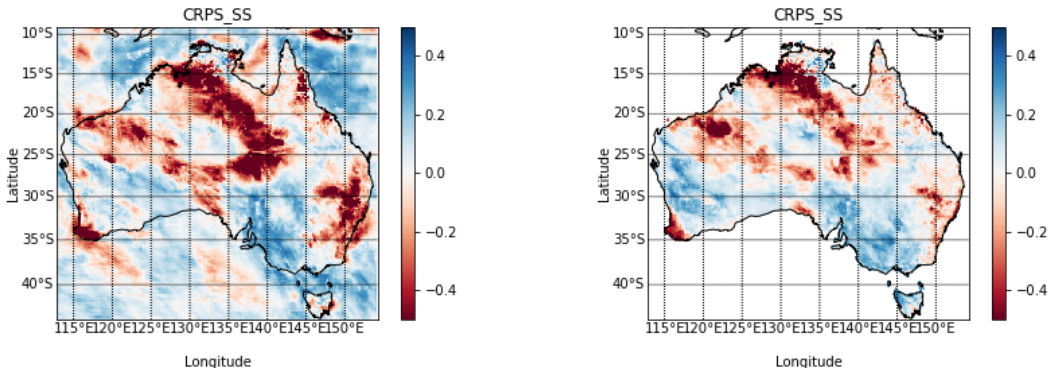
Fig. 9: CRPS Skill Score map with the lead time 6 days averaged across 48 initialisation dates in 2012



(a) BI (b) QM

(c) VDSR (d) VDSD

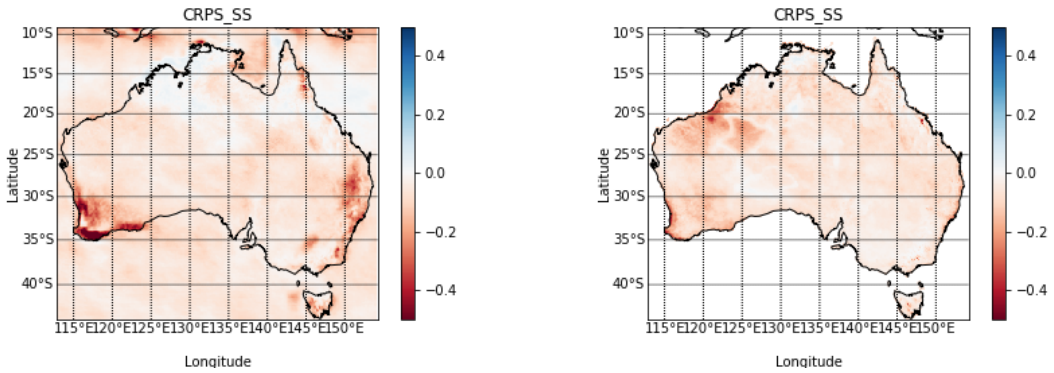
Fig. 10: Average CRPS skill score for lead time 0 to 216 days across Australia for forecasts made in 2012



(a) BI (b) QM

(c) VDSR (d) VDSD

Fig. 11: CRPS Skill Score map with the lead time 6 days averaged across 48 initialisation dates in 2010



(a) BI (b) QM

(c) VDSR (d) VDSD

Fig. 12: Average CRPS skill score for lead time 0 to 216 days across Australia for forecasts made in 2010