Downscaling Long Lead Time Daily Rainfall Ensemble Forecasts through Deep Learning

Huidong Jin[1*], Weifan Jiang[1,2], Minzhe Chen[2], Ming Li[3], K. Shuvo Bakar[1,4] and Quanxi Shao[3]

[1*]Data61, CSIRO, North Science Road, Acton, 2601, ACT, Australia.
[2]CECS, the Australian National University, North Road, Acton, 2601, ACT, Australia.
[3]Data61, CSIRO, PO BOX 1130, Bentley WA, 6102, WA, Australia.
[4]Faculty of Medicine and Health, Sydney University, Science Rd, Camperdown, 2050, NSW, Australia.

*Corresponding author(s). E-mail(s): warren.jin@csiro.au;
Contributing authors: u6683698@alumni.anu.edu.au;
chenminzhe2020@163.com; Ming.Li@csiro.au;
shuvo.bakar@sydney.edu.au; Quanxi.Shao@csiro.au;

Abstract

Skilful and high-resolution daily weather forecasts for upcoming seasons are of huge value to climate-sensitive sectors. General Circulation Models (GCM) provide routinely such long lead time ensemble forecasts, known as Seasonal Climate Forecasts (SCF), and require downscaling techniques to improve their spatial resolution and consistency with local observations. Traditional downscaling techniques, like Quantile Mapping (QM), learn relationship from historical observations and hindcasts, and are time-consuming or labour-intensive for operation. Almost all of deep-learning based downscaling techniques focus on simplified situations where low-resolution images match well with high-resolution images, which is not the case in SCFs. In this paper, after applying several image super-resolution models for downscaling rainfall, we choose Very Deep Super-Resolution (VDSR) as the best candidate, according to an overall ensemble forecast skill metric, Continuous Ranked Probability Score (CRPS). To improve forecast skills, we propose Very

Deep Statistical Downscaling (VDSD) model via further incorporating resolved variables such as geopotential height. 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 Error (MAE) and CRPS respectively, than VDSR and QM. The results also show that VDSD performs better than or comparably to climatology, a benchmark for long lead time climate forecasts. Deep learning techniques like VDSD are appealing for skilful and efficient SCF downscaling.

Keywords: Statistical downscaling; probabilistic/ensemble forecast; seasonal climate forecasts; image super-resolution; deep learning

1 Introduction

Seasonal Climate Forecasts (SCF) have great value to many socio-economic sectors such as agriculture, construction, mining, tourism, energy, and health (Manzanas, 2020; Merryfield et al., 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 schemes. SCFs have been estimated to contribute up to $8.40 and A$258 per hectare per year to USA and Australia agriculture respectively (Mjelde & Griffiths, 1998; Parton, Crean, & Hayman, 2019). For the whole of Australia, the 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 & Mahmood, 2016). To reach their full potential, SCFs provided should be timely and skilful in high spatial resolution to help these weather-sensitive sectors make evidence-based site-specific decisions (Li & Jin, 2020; Schepen, Everingham, & Wang, 2020).

After three decades of development, SCFs based on General Circulation Models (GCMs) have moved beyond the research realm and are routinely produced by climate centres or service agents around the world (Hudson et al., 2017; Johnson et al., 2019; Merryfield et al., 2020; Saha et al., 2014). These state-of-the-art GCMs couple together with physics-based models of ocean, atmosphere, land surface and sea-ice. They can capture synoptic-scale climate dynamics at grids of horizontal/spatial resolutions commonly around 100-200 km (Johnson et al., 2019; Liu, Ganguly, & Dy, 2020; Ratnam, Doi, & Behera, 2017; Vandal et al., 2017). These physical models also incorporate hundreds of semi-empirical relationships to approximate processes such as convection and cloud formation that are too fine for the models to resolve (Manzanas, 2020;

Vandal et al., 2017). Unfortunately, these empirical relationships may be ill-constrained. Furthermore, limited by computational resources, coarse spatial resolution and simplified nature of GCMs often make the produced forecasts inconsistent with observed weather, especially for precipitation or at 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 SCF from Australia's Bureau of Meteorology (BoM) has 11 ensemble members for each initialisation date and its United States counterpart has 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, Manzanas, & Gutiérrez, 2020; Kusunose & Mahmood, 2016; Schepen et al., 2020). The barriers may be overcome via downscaling techniques which generate more skilful and localised forecasts by making use of weather observations, and sometimes other localised information (Bettolli et al., 2021; Maraun & Widmann, 2018).

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 are a large number of downscaling techniques developed, including dynamical downscaling (Manzanas, 2020; Ratnam et al., 2017; Thatcher & McGregor, 2009), statistical downscaling (Maraun & Widmann, 2018), and the recent development of deep-learning-based downscaling. Dynamical downscaling uses a physics-based climate model, forced by boundary conditions from a GCM, to simulate atmospherical 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, quantify and adjust the uncertainty spread, and restore local daily climate variability of GCM forecasts (Ahmadalipour, Moradkhani, & Rana, 2018; Maraun & Widmann, 2018). 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 (Li & Jin, 2020; Michelangeli, Vrac, & Loukos, 2009). 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 varies, can be substantial or almost nothing, depending on locations and seasons (Ahmadalipour et al., 2018; Li & Jin, 2020; Manzanas, Lucero, Weisheimer, & Gutierrez, 2018; Schepen et al., 2020). 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 abilities to capture important 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), has 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 (Luo, Li, & Ye, 2021; Shi et al., 2017; Xingjian et al., 2015), which predict rainfall intensity in a local region over a relatively short period time, and precipitation parameterisations from GCMs (Pan, Hsu, AghaKouchak, & Sorooshian, 2019). More related to this study, several downscaling techniques have been developed based on Single Image Super-Resolution (SISR) techniques since (Vandal et al., 2017). For example, for long-term climate projection, Rodrigues, Oliveira, Cunha, and Netto (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. Three layers are relatively shallow. Super-Resolution Deep Residual Network (SRDRN) was proposed based on a deep convolutional neural network with residual blocks and batch normalisation for downscaling daily precipitation and temperature (F. Wang, Tian, Lowe, Kalin, & Lehrter, 2021). It leaves behind the bias-correction required in downscaling (F. Wang et al., 2021). 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, which have different characteristics from daily precipitation. These pioneering downscaling techniques have varying success.

Downscaling climate forecasts 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 (Z. Wang, Chen, & Hoi, 2020). Most deep-learning-based downscaling techniques focused on a simplified situation with a single data source (Liu et al., 2020; Rodrigues et al., 2018; Vandal et al., 2017; F. Wang et al., 2021).
2. Bias and displacement in space or time are common in climate forecasts, especially for precipitation, due to the inherent complexity of climate modelling. To mitigate these issues, multiple possible forecast trajectories are provided as a practical standard for short or long lead time forecasts (Hudson et al., 2017; Johnson et al., 2019; Merryfield et al., 2020). 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, Gneiting, Berrocal, & Johnson,

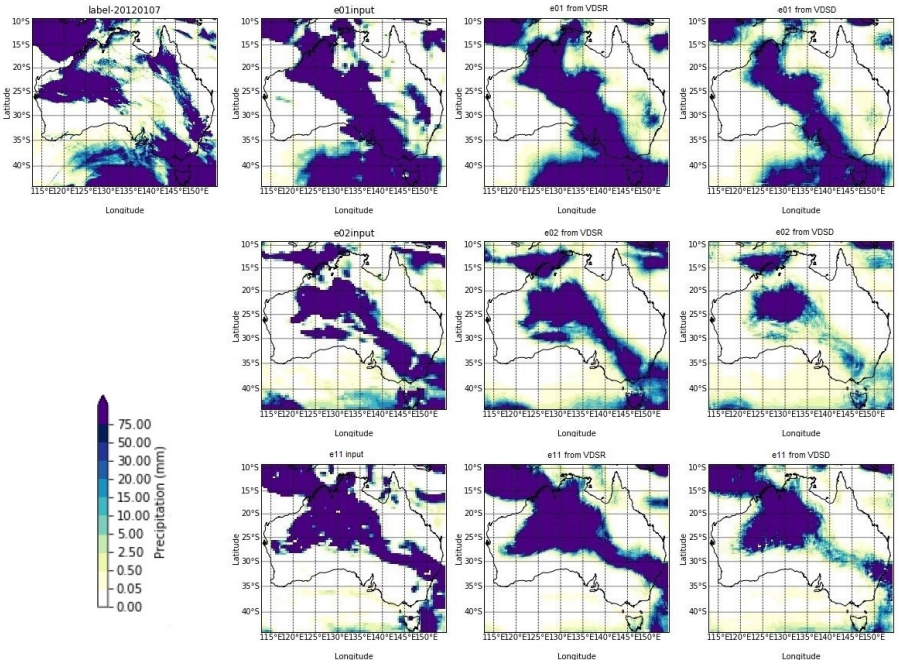


Fig. 1: Reanalysis data and daily rainfall ensemble forecasts for 7 Jan 2012 with a lead time of six-days for the forecasts made on 1 Jan 2012. Images in the four columns are the high-resolution image from BARRA reanalysis data, ensemble member forecasts from ACCESS-S1 after bicubic interpolation, and downscaled results of VDSR and VDSD respectively. Only the 1st, 2nd, and 11th members are illustrated.

2006; Kusunose & Mahmood, 2016; Li & Jin, 2020). The latter is predominant in climate communities (Ferro, Richardson, & Weigel, 2008; Grimit et al., 2006; Schepen et al., 2020) but, as far as we know, has never been used in deep learning downscaling development.

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

To address these differences, for downscaling long lead time daily precipitation forecasts in Australia, we choose Very Deep Super-Resolution (VDSR) (Kim, Kwon Lee, & Mu Lee, 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 (Ferro et al., 2008; Grimit et al., 2006; Li & Jin, 2020; 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 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 present climate data in Section 2. We brief SISR models, and select three deep-learning based image super-resolution models and then propose and finalise the new downscaling model VDSD in Section 3. Cross-validation and comparison results are given in Section 4. We conclude the paper in Section 5 with discussions on further developments of VDSD for possible operation use.

2 Data and Pre-processing

2.1 ACCESS-S1 retrospective forecast raw and calibrated 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) (Bureau National Operations Centre, 2019; Hudson et al., 2017), which 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 large ensembles. The resolution of the atmospheric model is raised to 0.6°, (nearly 60km×60km), as the Stochastic Kinetic Energy Backscatter scheme (Bowler, Arribas, Beare, Mylne, & Shutts, 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). 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 a lead time of 0-216 days, and 11 ensemble members. Each forecast member provides a full description of the evolution of weather for the upcoming 217 days, and collectively these ensemble forecasts indicate the likelihood of a range of future weather scenarios. Daily precipitation data from

[1] We call these 'forecast' hereafter in the paper for simplicity.
[2] http://poama.bom.gov.au/general/hindcast_data.html.

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 of 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. Bilinear interpolation is performed using linear interpolation first in one direction from one resolution to another resolution, and then again in the other 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 probability distributions (CDFs, aka quantile functions) of raw forecasts x_f and observations x_o respectively (Maraun & 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 (Bureau National Operations Centre, 2019; Li & Jin, 2020).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 hereafter.

2.2 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 (Acharya, Nathan, Wang, Su, & Eizenberg, 2019). Its spatial resolution of 0.12° (i.e., around 12km×12km), is realised in the whole region of Australia and New Zealand. BARRA uses 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 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

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

of the daily aggregation is based on the same-day definition of ACCESS-S1 data.

2.3 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 all the Australian landmass (see, Fig. 1). As preprocessing, 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, and 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 super-resolution, 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 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 for training.

3 Deep Learning for Downscaling Ensemble Forecasts

To develop deep learning techniques for downscaling daily rainfall ensemble forecasts, we first briefly describe Single Image Super-Resolution (SISR) techniques and several successful deep learning models.

3.1 Image Super-resolution and Deep learning

We model our downscaling problem as SISR where the GCM raw forecast and BARRA reanalysis daily data are treated as low- and high- resolution images respectively. SISR is to recover 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 a degradation mapping function; H is a high-resolution image corresponding to L; γ is the parameters of the degradation mapping function (Z. Wang et al., 2020). Most super-resolution data sets are in fact obtained by aggregation or degradation mapping from high-resolution images (Z. Wang et al., 2020). 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 $\mathcal{F}$ and its parameter θ.

The simplest SISR techniques are spatial interpolation, such as nearest-neighbour interpolation, bilinear interpolation, and Bicubic Interpolation (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 artifacts. BI is often considered to be the baseline for spatial downscaling of precipitation fields (Vandal et al., 2017).

Since Dong, Loy, He, and Tang (2014) first introduced Super-Resolution CNN (SRCNN) to the SISR task, deep-learning-based SISR techniques have been widely developed and achieved great improvements in terms of image or perceptual quality. Most of them are based on Convolutional Neural Networks (CNN) (Liu et al., 2020). As surveyed by Z. Wang et al. (2020), these SISR models use several 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, Tian, Kong, Zhong, & Fu, 2018), and attention mechanism in Residual Channel Attention Network (RCAN) (Zhang, Li, et al., 2018). RCAN achieve state-of-the-art in terms of image quality measured by Peak Signal-to-Noise Ratio (PSNR) (Z. Wang et al., 2020). To generate more realistic images, encoder-decoder network or generative adversarial network (GAN), are used in Super-Resolution GAN (SRGAN) (Ledig et al., 2017) and Enhanced SRGAN (ESRGAN) (X. Wang et al., 2018), and multiple semantic information used in (Rad et al., 2019). ESRGAN outperforms various models in terms of perceptual quality.

We develop our deep learning technique for ensemble SCFs in two phases. First, we select several deep learning techniques based on their superior SISR performance in literature (Kim et al., 2016; X. Wang et al., 2018; Zhang, Li, et al., 2018), and then train them to generate a high-resolution precipitation image from each low-resolution 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 (Section 3.2). Second, based on the selected very deep learning structure, we incorporate other resolved climate variables and propose VDSD to enhance its downscaling performance. More details for these two steps are described in the following two subsections respectively.

3.2 Model selection for downscaling ensemble forecasts

GCMs are always a substantial simplification of the real-world climate system, which is complex and high-dimensional (Maraun & 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 operational 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 a lead time of 6 days. All the ensemble members target the same date, Jan 7 Dec 2012, and share the same target images. The forecast accuracy metrics such as Mean Absolute Error (MAE), $\frac{\sum_{j,k}\left\|\hat{y}_{j,k}^{(i,l,e)} - y_{j,k}^{(d)}\right\|}{\sum_{j,k} 1}$ [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 in 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 (Grimit et al., 2006; Hersbach, 2000). It is defined as

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

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 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 correspond across ensembles. 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 (Basso & Liu, 2019; Jin, Li, Hopwood, Hochman, & Bakar, 2022; Schepen et al., 2020). 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 $Z^{(i,l,e)} \in \mathcal{R}^{m_0 \times n_0 \times p}$

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

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 $Y^{(d)}$,

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

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

$$\overline{CRPS} = \frac{\sum\limits_{i,l,j,k} w_{j,k}^{(i,l)} CRPS\left(\hat{F}_{j,k}^{(i,l)}, y_{j,k}^{(d)} \right)}{\sum\limits_{i,l,j,k} w_{j,k}^{(i,l)}} \tag{4}$$

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 $w_{j,k}^{(i,l)} \equiv 1$ for this study for simplicity.

To determine 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 an 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 are 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, Li, et al., 2018), and ESRGAN (X. Wang et al., 2018). They are chosen because of their outstanding performance on SISR (X. Wang et al., 2018; Z. Wang et al., 2020). The training is based on image super-resolution, and our deep learning-based problem becomes:

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

where $\mathcal{L}$ is the loss function, calculating an error between high-resolution images $Y^{(d)}$ and super-resolution images $\hat{Y}^{(i,l,e)}$ (output from Eq 3); λ is a tradeoff parameter and $\Phi(\theta)$ is the regularisation term.

On the separate validation data sets, the average CRPS skill score of trained ESRGAN across 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 forecast skill of trained RCAN is not as good as QM in the validation. VDSR is much faster to converge and outperforms QM in terms of average CRPS. We select the VDSR model for further

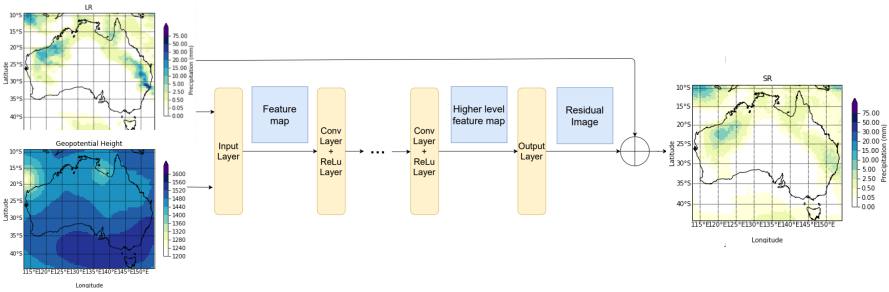


Fig. 2: The structure of the VDSD model, modified from VDSR (Kim et al., 2016), where $\oplus$ represents element-wise matrix addition with input precipitation image, orange blocks are layers of the neural network, blue rectangles are feature maps, input and output daily rainfall data images are on the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

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 CRPS decreases from 8 to 18 layers, and after that, did not change that much in the held-out validation. We stick with 18 layers as recommended by Kim et al. (2016) for SISR.

3.3 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 (Baño-Medina et al., 2020; Pan et al., 2019). 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 (Baño-Medina et al., 2020; Merryfield et al., 2020; Pan et al., 2019). The overall structure of finalised VDSD is illustrated in Fig 2, where Geopotential Height (ZG) at 850 hPa is used as the additional input. This ZG 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, such as ZG as input like in Eq 3. 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 2.3). 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 target image. VDSD maintains the residual learning which has been widely demonstrated in robust and speedy training in SISR (Kim et al., 2016; Z. Wang et al., 2020).

As shown in Fig 2, 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

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

where M_0 is the first level feature map 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 kernels and produces 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 images. 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

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

where M_t represents the tth level feature map, and 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 of Australia on the separate validation data set to finalise the structure of VDSD. We test two different types of variants of VDSD. (1) One adds extra input images for downscaling. We try four climate variables from ACCESS-S1 and their combinations. They are ZG (at 850 hPa), daily maximum temperature, daily minimum temperature, and sea level pressure. Our results show that adding more climate

variables than ZG rarely reduces CRPS, and sometimes deteriorate its ensemble forecast skills. (2) The other is to try two different loss functions in Eq 5, 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, $\mathcal{L}\left(\hat{Y}^{(i,l,e)}, Y^{(d)}\right) = \sqrt{\frac{\sum_{k,j}\left(\hat{y}_{k,j}^{(i,l,e)} - y_{k,j}^{(d)}\right)^2}{\sum_{k,j} 1}}$. L1 gives us better validation CRPS, and would be used in our final VDSD model. Note that in VDSR, L2 is preferred.

4 Results and Comparison

To illustrate the downscaling performance of VDSR and VDSD that we've finalised, we used the last three 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 for 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. In total, around 1152 daily precipitation images from BARRA were used in prediction performance validation and comparison.

4.1 Performance Metrics for Forecasts

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 & 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 we discuss in Section 3.2, the average CRPS of ensemble forecasts for each grid point (k, j) 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 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 with 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 based on deep learning are often assessed by MAE (Liu et al., 2020; Z. 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 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 exclude locations on the ocean as the QM model focused on the Australian continent.

4.2 Results for forecasts made 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 spatially-interpolated ACCESS-S1 raw forecasts and often has more areas with precipitation. Downscaled results of VDSD follow precipitation shapes of the raw forecasts while more likely reduce 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.

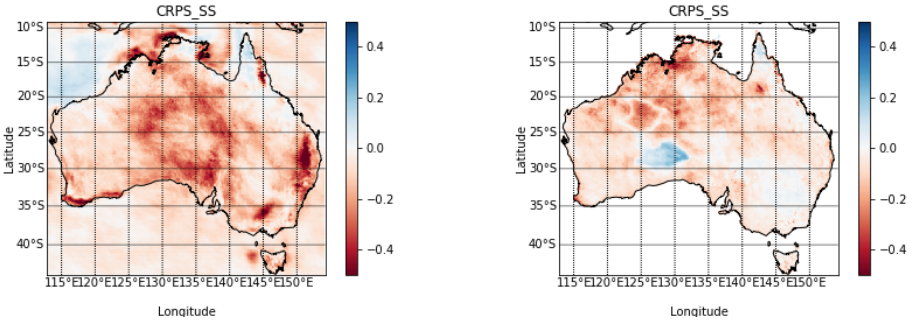
Averaging across 48 initialisation days in 2012, we calculate its average CRPS skill score for each grid point and lead time. Fig 3 illustrates mean CRPS skill scores across the whole of Australia by the four downscaling models along with lead time up to 216 days. VDSR has the highest scores in the first three lead times, and then VDSD becomes the best of the four models for almost all 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. S1 [5]. For most locations on the Australian continent (except north-western Australia, the 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, VDSD has positive CRPS skill scores for most locations in Australia, while its three counterparts have negative skills for most locations (Fig. S2).

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

Fig. 3: Average CRPS Skill Scores across the whole Australian land for the forecasts made on 48 different initialisation dates in 2012.

That means only VDSD is better than the climatology for most locations. 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 and QM than that of VSDR and 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 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}. VDSD has positive skill scores for most locations on the Australian continent, with an mean of around 2.76×10^{-2}. VDSD still has negative



(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 made on 48 different initialisation dates in 2012

skill scores along the eastern coastline, north-western parts of Australia, and Tasmania.

Fig 5 illustrates the MAE skill scores of four downscaling techniques along lead times. Except for the first six lead times, BI has negative skill scores, indicating that the ACCESS-S1 raw forecasts have limited skill. QM often has positive MAE skill scores. Both deep learning models, VDSR and VDSD have substantial improvements for all the different lead times. Averaging across these 217 different lead times, the MAE skill scores of VDSD, VDSR, QM and BI are around 0.38, 0.19, 0.02, and -0.13 respectively.

4.3 Results for forecasts made in 2010

Fig 6 illustrates the average CRPS skill scores along with lead time based on forecasts made on 48 different initialisation days in 2010. For most 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 in most locations in Australia. In comparison,

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

QM and BI have negative CRPS skill scores in many locations (Fig. S3). 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 of VDSD and VDSR across the 217 different lead times are often positive or close to zero for most locations on Australian land (Figs S4d and S4c), and both QM and BI are normally in the negative domain (Figs S4b and S4a). 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 lead to a few percentage points lower on CRPS skill score (Ferro et al., 2008; Li & 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 in Australia, both VDSR and

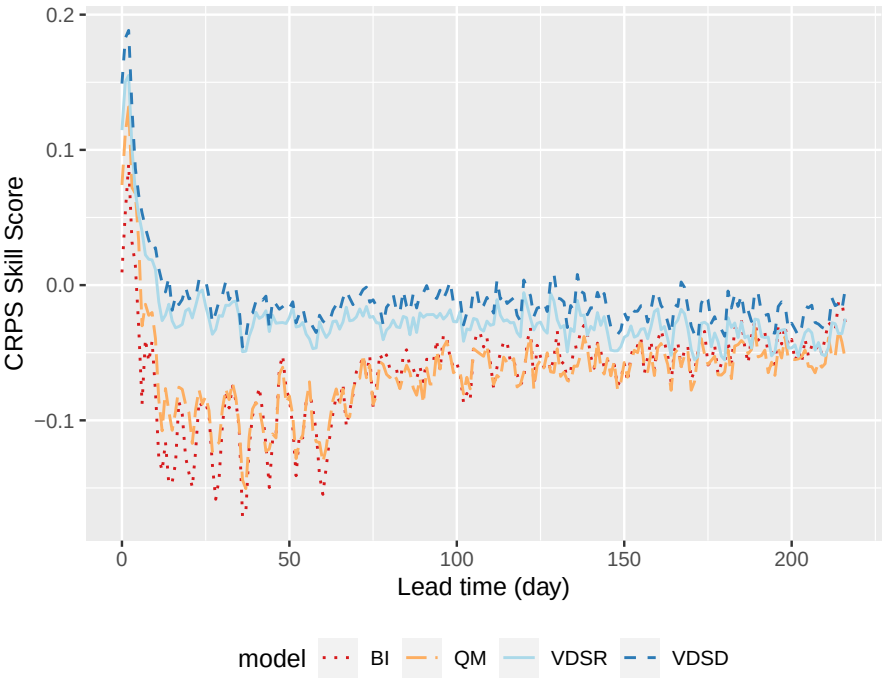


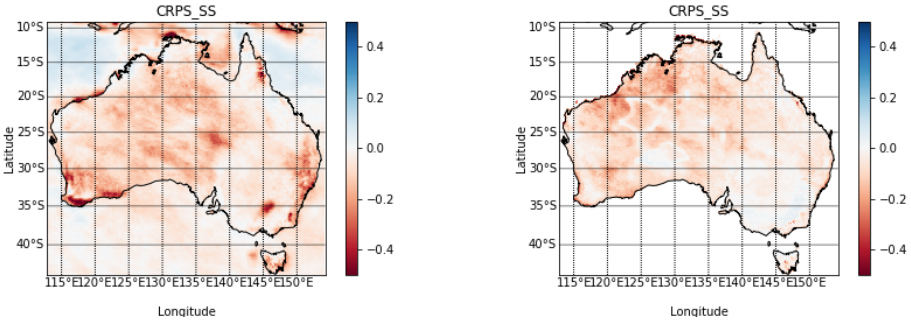
Fig. 6: Average CRPS skill scores across the whole Australian land for forecasts made on 48 initialisation dates in 2010.

VDSD have between -0.1 and 0.1 CRPS skill scores while 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 conclude VDSD is comparable with climatology in terms of both forecast accuracy and ensemble forecast skill for SCFs made on the 48 initialisation dates in 2010.

4.4 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 observed the objective function stabilised. The learning rate was 1.0×10^{-4}, and relatively small as the network is very deep, and a large learning rate may cause the vanishing/exploding gradient problem (Bengio, Simard, & Frasconi, 1994). The



(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

optimisation method is stochastic gradient descent with a momentum of 0.9. Our implementation is written 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 downscaling and validation was done on a normal PC with a mid-range GPU GeForce RTX 2070.

Table 2 lists the average computation time required for both training and operation where 11 ensemble members for 217 days 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 on the normal 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 & Jin, 2020) needed about one hour for training for a station or grid point and took 0.46 seconds for an operation forecast. It

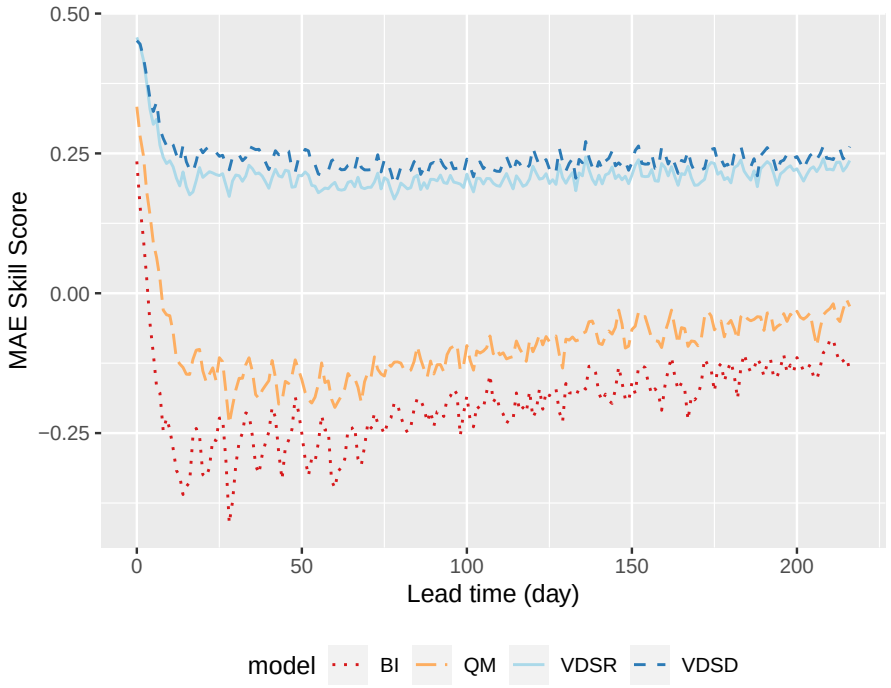


Fig. 8: Average MAE skill scores across Australia for precipitation forecasts made on 48 initialisation dates.

	Gadi in NCI.org.au	PC for downscaling forecasts
CPU	36× Intel®Xeon™ Platinum 8268	1× Intel®Core™ 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

Table 1: Hardware configuration used in the experiments

would take around 15.1 hours to downscale rainfall for the whole Australia on a normal PC. The dynamic downscaling model, Conformal Cubic Atmospheric Model (CCAM) (Thatcher & McGregor, 2009), doesn'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 downscaling seasonal rainfall forecasts.

Table 2: 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
VDSR	12.12	0.08
VDSD	16.76	0.56

5 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 to generate and select high-level features automatically, 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 has 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. Thus, deep learning models, especially the proposed VDSD, have demonstrated their potential for possible operational use in the 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 [6]. The La Niña event peak in 2012 is much weaker, and made 2012 relatively easier to forecast. That means the training data for models to test on 2010 and 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 in 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 percentage points higher on CRPS (Ferro et al., 2008). Therefore,

[6] http://www.bom.gov.au/climate/history/enso/. Accessed on 20 Jan 2022.

although the CRPS skill score is negative on average, the proposed VDSD is thought to be comparable with climatology. It is still appealing because it needs less downscaling operation time than QM.

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 & 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 not be 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 & Jin, 2020) can further improve long lead time forecasts. As the spatial and cross-variable relationships are not necessarily 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, Jin, & Shao, 2021). VDSD only downscales to 12km, which should be further increased for real-world applications. Deep learning still demands a lot of time for model development, finalisation and training for downscaling, and we will investigate suitable models for other climate variables. For a fair comparison and training time reduction, we have only included the forecasts with lead time less than seven days in the training data. That may lead to putting more emphasis on low-resolution precipitation and less on correcting inherent biases of GCM's forecasts. A good tradeoff between bias correction and resolution improvement is also subject to future work.

References

Acharya, S.C., Nathan, R., Wang, Q.J., Su, C.-H., Eizenberg, N. (2019). An evaluation of daily precipitation from a regional atmospheric reanalysis over Australia. *Hydrology and Earth System Sciences*, *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. *Climate Dynamics*, *50*(1-2), 717-733. https://doi.org/10.1007/s00382-017-3639-4

Baño-Medina, J., Manzanas, R., Gutiérrez, J.M. (2020). Configuration and intercomparison of deep learning neural models for statistical downscaling. *Geoscientific Model Development*, *13*(4), 2109–2124.

Basso, B., & Liu, L. (2019). Seasonal crop yield forecast: Methods, applications, and accuracies. In *Advances in agronomy* (Vol. 154, p. 201-255).

Elsevier. 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 transactions on neural networks*, *5*(2), 157–166.

Bettolli, M., Solman, S., Da Rocha, R., Llopart, M., Gutierrez, J., Fernández, J., ... others (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. *Climate Dynamics*, *56*(5), 1589–1608.

Bowler, N.E., Arribas, A., Beare, S.E., Mylne, K.R., Shutts, G.J. (2009). 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*, *135*(640), 767–776.

Bureau National Operations Centre (2019, Sep). *Operational implementation of ACCESS-S1 forecast post processing* (Tech. Rep. No. 124). Melbourne VIC 3001: Bureau of Meteorology.

Davies, T., Cullen, M.J., Malcolm, A.J., Mawson, M., Staniforth, A., White, A., Wood, N. (2005). A new dynamical core for the met office's global and regional modelling of the atmosphere. *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, *131*(608), 1759–1782.

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

Ferro, C.A., Richardson, D.S., Weigel, A.P. (2008). On the effect of ensemble size on the discrete and continuous ranked probability scores. *Meteorological Applications*, *15*(1), 19-24.

Grimit, E.P., Gneiting, T., Berrocal, V.J., Johnson, N.A. (2006). The continuous ranked probability score for circular variables and its application to mesoscale forecast ensemble verification. *Quarterly Journal of the Royal Meteorological Society*, *132*(621C), 2925–2942.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, *15*(5), 559-570. https://doi.org/10.1175/1520-0434(2000)015⟨0559:DOTCRP⟩2.0.CO;2

Hudson, D., Alves, O., Hendon, H.H., Lim, E.-P., Liu, G., Luo, J.-J., ... et al (2017). ACCESS-S1 the new bureau of meteorology multi-week to seasonal prediction system. *Journal of Southern Hemisphere Earth Systems Science*, *67*(3), 132–159. https://doi.org/10.1071/ES17009

Jin, H., Li, M., Hopwood, G., Hochman, Z., Bakar, K.S. (2022). Improving early-season wheat yield forecasts driven by probabilistic seasonal climate forecasts. *Agricultural and Forest Meteorology*, *315*, 108832. https://doi.org/10.1016/j.agrformet.2022.108832

Johnson, S.J., Stockdale, T.N., Ferranti, L., Balmaseda, M.A., Molteni, F., Magnusson, L., ... Monge-Sanz, B.M. (2019). SEAS5: the new ECMWF seasonal forecast system. *Geoscientific Model Development*, *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. *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. *Agricultural Systems*, *146*, 103-110. https://doi.org/10.1016/j.agsy.2016.04.006

Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., ... Wang, Z. (2017). Photo-realistic single image super-resolution using a generative adversarial network. *Proceedings of the ieee conference on computer vision and pattern recognition* (p. 4681-4690).

Li, M., & Jin, H. (2020). Development of a postprocessing system of daily rainfall forecasts for seasonal crop prediction in Australia. *Theoretical and Applied Climatology*, *141*, 1331-1349. https://doi.org/10.1007/s00704-020-03268-3

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

Liu, Y., Ganguly, A.R., Dy, J. (2020). Climate downscaling using YNet: A deep convolutional network with skip connections and fusion. *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 3145–3153).

Luo, C., Li, X., Ye, Y. (2021). 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*, *14*, 843-857. 10.1109/JSTARS.2020.3040648

MacLachlan, C., Arribas, A., Peterson, K., Maidens, A., Fereday, D., Scaife, A., ... et al (2015). Global seasonal forecast system version 5 (glosea5): a high-resolution seasonal forecast system. *Quarterly Journal of the Royal Meteorological Society*, *141*(689), 1072–1084.

Manzanas, R. (2020). Assessment of model drifts in seasonal forecasting: Sensitivity to ensemble size and implications for bias correction. *Journal of Advances in Modeling Earth Systems*, *12*(3), e2019MS001751.

Manzanas, R., Lucero, A., Weisheimer, A., Gutierrez, J.M. (2018). Can bias correction and statistical downscaling methods improve the skill of seasonal precipitation forecasts? *Climate Dynamics*, *50*(3-4), 1161-1176. https://doi.org/10.1007/s00382-017-3668-z

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

Merryfield, W.J., Baehr, J., Batté, L., Becker, E.J., Butler, A.H., Coelho, C.A., ... et al (2020). Current and emerging developments in subseasonal to decadal prediction. *Bulletin of the American Meteorological Society*, *101*(6), E869–E896.

Michelangeli, P.A., Vrac, M., Loukos, H. (2009). Probabilistic downscaling approaches: Application to wind cumulative distribution functions. *Geophysical Research Letters*, *36*. ArtnL1170810.1029/2009gl038401

Mjelde, J.W., & Griffiths, J.F. (1998). A review of current evidence on climate forecasts and their economic effects in agriculture. *American Journal of Agricultural Economics*, *80*(5), 1089–1095.

Pan, B., Hsu, K., AghaKouchak, A., Sorooshian, S. (2019). Improving precipitation estimation using convolutional neural network. *Water Resources Research*, *55*(3), 2301-2321. https://doi.org/10.1029/2018WR024090

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

Rad, M.S., Bozorgtabar, B., Marti, U.-V., Basler, M., Ekenel, H.K., Thiran, J.-P. (2019). SROBB: Targeted perceptual loss for single image super-resolution. *Proceedings of the ieee/cvf international conference on computer vision* (pp. 2710–2719).

Ratnam, J., Doi, T., Behera, S.K. (2017). Dynamical downscaling of SINTEX-F2v CGCM seasonal retrospective austral summer forecasts over Australia. *Journal of Climate*, *30*(9), 3219–3235.

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., et al. (2019). Deep learning and process understanding for data-driven earth system science. *Nature*, *566*(7743), 195–204.

Rodrigues, E.R., Oliveira, I., Cunha, R., Netto, M. (2018). DeepDownscale: a deep learning strategy for high-resolution weather forecast. *2018 ieee 14th international conference on e-science (e-science)* (pp. 415–422).

Saha, S., Moorthi, S., Wu, X., Wang, J., Nadiga, S., Tripp, P., ... et al (2014). The NCEP climate forecast system version 2. *Journal of climate*, *27*(6), 2185–2208. https://doi.org/10.1175/JCLI-D-19-0230.1

Schepen, A., Everingham, Y., Wang, Q.J. (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. *Agricultural and Forest Meteorology*, *291*, 107991. https://doi.org/10.1016/j.agrformet.2020.107991

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

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

Thatcher, M., & McGregor, J.L. (2009). Using a scale-selective filter for dynamical downscaling with the conformal cubic atmospheric model. *Monthly Weather Review*, *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. Retrieved from http://www.climatekelpie.com.au/Files/MCV-CIE-report-Value-of-improved-forecasts-non-agriculture-2014.pdf (Accessed in Nov 2020)

Vandal, T., Kodra, E., Ganguly, S., Michaelis, A., Nemani, R., Ganguly, A.R. (2017). DeepSD: Generating high resolution climate change projections through single image super-resolution. *KDD'17* (p. 1663-1672). https://doi.org/10.1145/3097983.3098004

Wang, F., Tian, D., Lowe, L., Kalin, L., Lehrter, J. (2021). Deep learning for daily precipitation and temperature downscaling. *Water Resources Research*, *57*(4), e2020WR029308.

Wang, X., Yu, K., Wu, S., Gu, J., Liu, Y., Dong, C., ... Change Loy, C. (2018). ESRGAN: Enhanced super-resolution generative adversarial networks. *Proceedings of the european conference on computer vision (eccv) workshops* (p. 63-79).

Wang, Z., Chen, J., Hoi, S.C. (2020). Deep learning for image super-resolution: A survey. *IEEE transactions on pattern analysis and machine intelligence*, *43*(10), 3365–3387.

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

Zhang, Y., Li, K., Li, K., Wang, L., Zhong, B., Fu, Y. (2018). Image super-resolution using very deep residual channel attention networks. *Proceedings of the european conference on computer vision (eccv)* (pp. 286–301).

Zhang, Y., Tian, Y., Kong, Y., Zhong, B., Fu, Y. (2018). Residual dense network for image super-resolution. *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2472–2481).

Acknowledgments. The authors would like to thank multiple CSIRO colleagues and BoM researchers, including Andrew Moore, Marcus Thatcher, Yanchang Zhao, Robert Smalley, and Morwenna Griffiths for their discussion and helps for 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.

Statements and Declarations

Funding

This work was partially funded by the CSIRO Digiscape Future Science Platform and the CAS-CSIRO Partnership program.

Competing interests

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

Author contributions

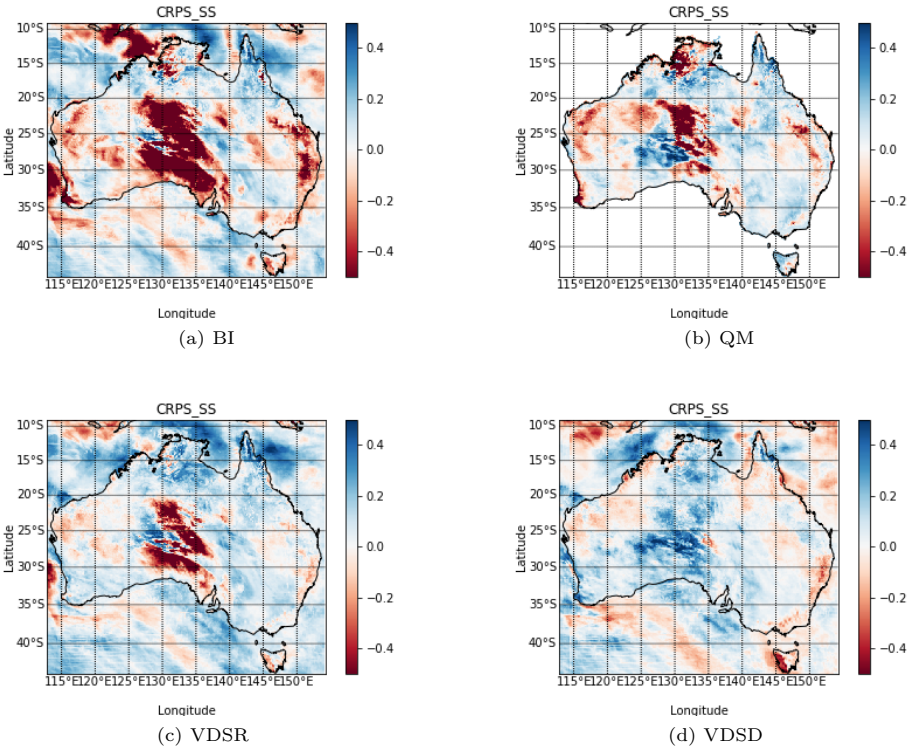
All authors contributed to the study's conception and design. Material preparation, data collection and analyses were performed by Huidong Jin, Weifan Jiang, and Minzhe Chen. The first draft of the manuscript was prepared by Huidong Jin and Weifan Jiang and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Data availability

All the data sets used in this study stored on NCI.org.au, which are accessible after applying for the following research project membership. ACCESS-S1 raw data and their calibrated versions are stored under the project of "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 of "High Resolution Seasonal Climate Forecast".

The source codes and some downscaled data are available in the repository https://github.com/JiangWeiFanAI/HRSCF.

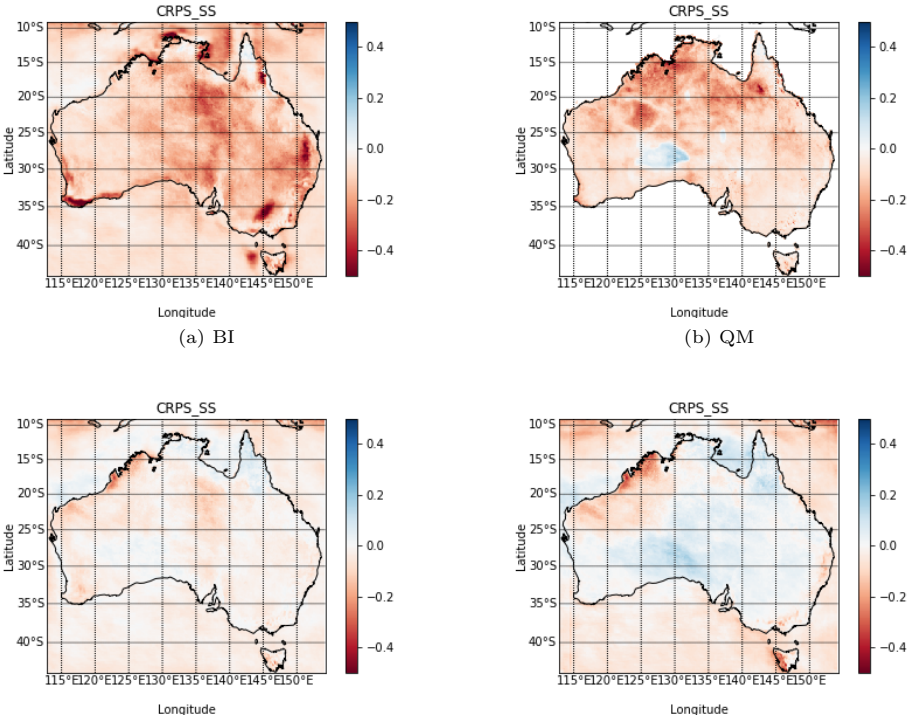
Supplementary information. Some figures are provided below as supplementary information for this manuscript.



(a) BI (b) QM

(c) VDSR (d) VDSD

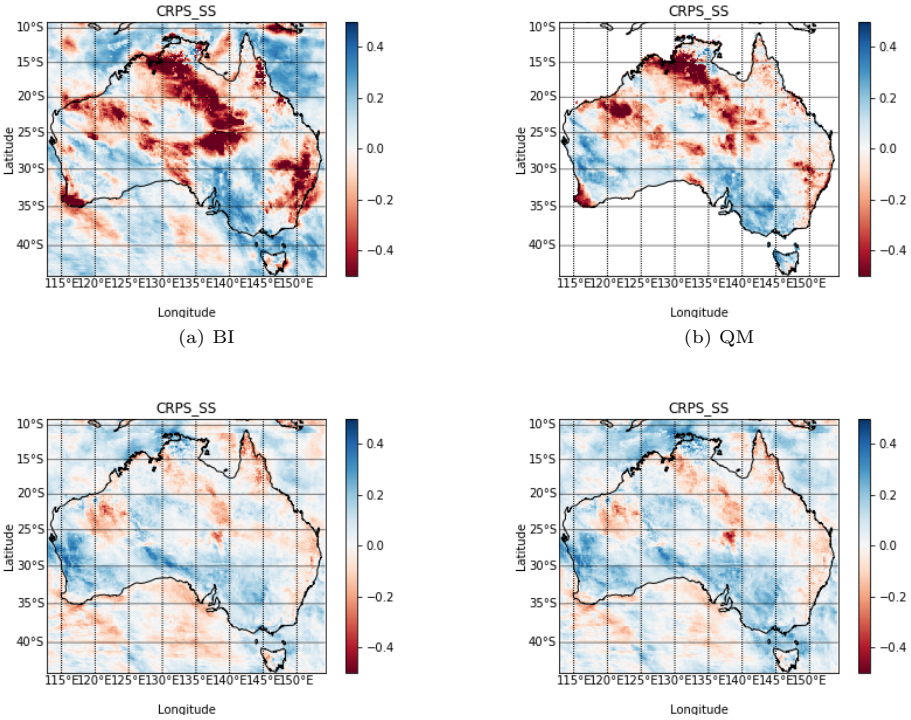
Fig. S1: CRPS skill score visualisation with the lead time of 6 days averaged across 48 initialisation dates in 2012



(a) BI (b) QM

(c) VDSR (d) VDSD

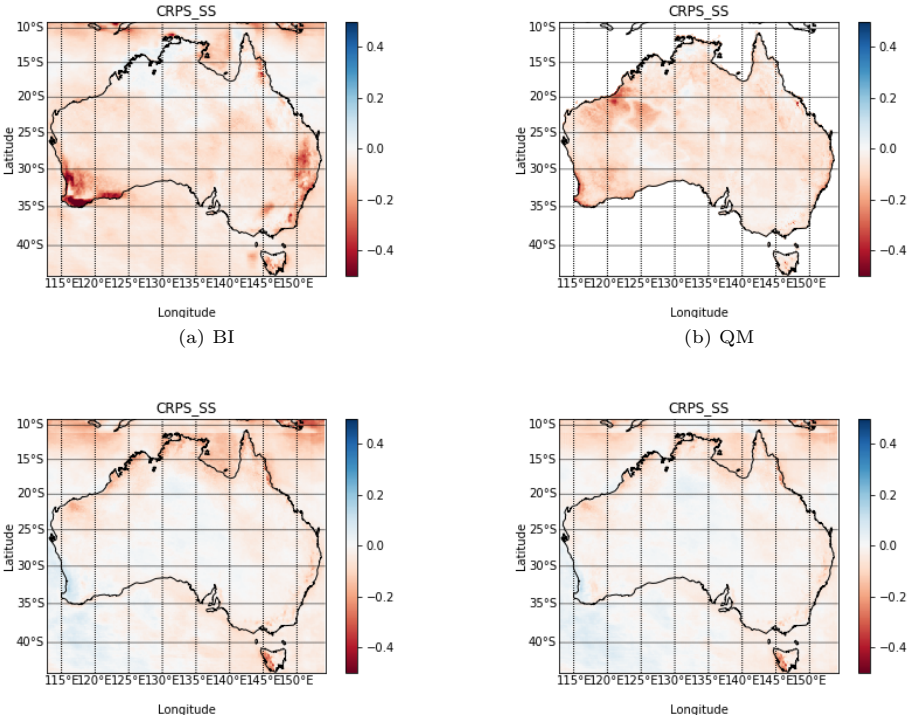
Fig. S2: Average CRPS skill score for lead time of 0 to 216 days across Australia for forecasts made in 2012



(a) BI (b) QM

(c) VDSR (d) VDSD

Fig. S3: CRPS skill score visualisation with the lead time of 6 days averaged across 48 initialisation dates in 2010



(a) BI (b) QM

(c) VDSR (d) VDSD

Fig. S4: Average CRPS skill score for lead time of 0 to 216 days across Australia for forecasts made in 2010