Applying Deep Learning for Operational Long Lead Time Probabilistic Daily Rainfall Forecasts*

Huidong Jin
Canberra ACT 2601 Australia

Weifan Jiang
ANU and CSIRO Data61

Minzhe Chen
ANU

Ming Li
BENTLEY WA 6102 Australia

K. Shuvo Bakar
Canberra ACT 2601 Australia

***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. Global Climate Models (GCM) are routinely providing long lead time ensemble climate forecasts while require downscaling techniques to improve their spatial resolution and consistency with local observations to be skilful. Traditional downscaling techniques, which use historical climate observations to learn a low resolution to finer resolution mapping, have limited skill improvement and often time-consuming for operation. Downscaling techniques based image superresolution have successfully been developed while most of them focused on simplified situations where low resolution images match reasonably well high resolution ones, which are not the case in operational long lead time daily rainfall forecasts. After applying several deep learning models for downscaling problem, we choose Very Deep SuperResolution (VDSR) as the most suitable candidate, according to an ensemble forecast skill metric, Continuous Ranked Probability Score (CRPS). We then propose Very Deep Statistical Downscaling (VDSD) model via incorporating resolved climate variables such as geopotential height, where extra features extracted can improve downscaling performance. Both VDSR and VDSD are tested on real-world applications for downscaling 60km ACCESS-S1 rainfall forecasts to 12km BARRA rainfall data with up to 216 days lead time for whole Australia. Leave-one-year-out cross-validation results illustrate VDSD has higher forecast accuracy and skill, measured by Mean Absolute Error (MAE) and CRPS respectively, than VDSR and traditional downscaling techniques. The results further show that VDSD performs better than or similarly to climatology, a benchmark for long lead time climate forecast. VDSD are lightweight to run on modest computer systems. It has potential for downscaling operation after several further developments.**

***Index Terms*—Statistical downscaling, probabilistic forecast, image superresolution, seasonal climate forecasts**

I. Introduction

Skilful Seasonal Climate Forecasts (SCF) have huge potential to improve productivity and profitability in weather-sensitive sectors such as agriculture, energy, mining and construction [1]. For example, daily rainfall forecasts for upcoming multiple months can benefit the whole agriculture value chains, such as helping farmers adapt their farming planning and management, and insurers and traders adjust their pricing scheme. For whole Australia, potential annual value added from skilful SCFs would be around $1.6 billion for the agricultural sector and $192 million for the construction sector [2]. To fulfill these potential values, SCFs should be provided timely and skilful in high spatial resolution so as to help weather-sensitive sectors make evidence-based site-specific decisions [**?**], [**?**].

After the last three decades development, SCFs using global climate models (GCMs) has moved beyond the research realm and are routinely produced by climate forecast centres around the world [1], [3], [4]. GCMs couple together physics-based models of ocean, atmosphere, land surface and sea-ice. GCMs can capture synoptic scale climate dynamics. GCMs are gridded with spatial resolutions commonly around 100 km [5], [6]. 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 [5]. 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. To improve forecast skills and quantifying uncertainty, ensemble forecasts, i.e., multiple simulations of a single model each with different initial conditions, The coarse spatial resolution and limited skills in representing local climate characteristics of GCMs circumvent the direct use of outputs in weather-sensitive applications [**?**], [7]. The barriers are normally bridged via downscaling techniques which generates more skilful and localised forecasts by making use of localised observations.

Downscaling is generally difficult and computationally expensive because of the complex nature of spatial-temporal structure of high resolution climate variables, especially for precipitation. There is a large body of downscaling technique development, including dynamical downscaling, statistical downscaling, and recent development on deep learning-based downscaling [8]. Comparisons between traditional statistical and dynamical downscaling suggest that neither group of methods is clearly superior, however in practice computationally cheaper statistical methods are widely used [7]. The skill improvement of statistical downscaling for long lead time daily forecasts can be substantial or almost nothing, depending on locations and seasons [**?**], [**?**]. The inadequacy of these statistical downscaling techniques may stem from the pre-engineered features and relationships prior to the modeling process, rarely exploiting their spatio-temporal dependencies

The work was funded by CSIRO digiscape future science platform.

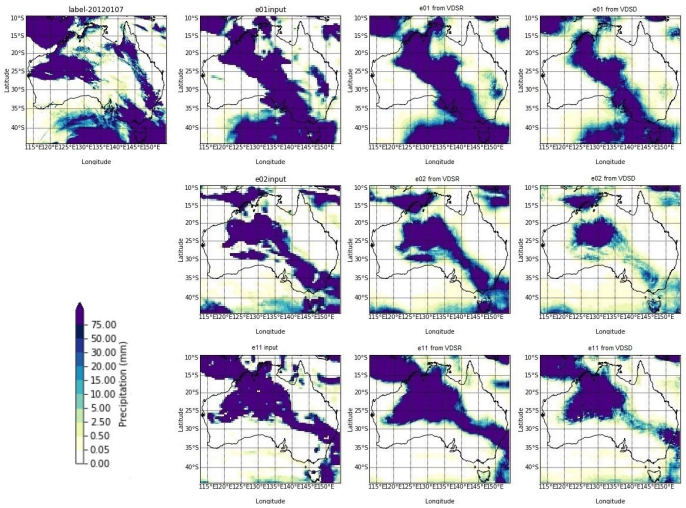

Fig. 1: Label image and ensemble rainfall forecasts for 7 Jan 2012 with six-day lead time for forecasts made on 1 Jan 2012. Images in the four columns are the high resolution label image from BARRA, ensemble member forecasts from ACCESS-S1 after bicubic interpolation, downscaled results of VDSR and of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

exhaustively. This limits their ability to capture important information beyond prior knowledge [6], [7]. Automatic feature extraction and selection integrated in 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 [9]. Deep learning has successfully been used in precipitation nowcasting [10], [11] and precipitation parameterisations from GCMs [12]. More related to our objective, downscaling techniques are developed based on Single Image Super-Resolution (SISR) techniques [6], [13]. To reuse deep learning based SISR techniques, most works over-simplify downscaling problems by taking the upscaled observation rather than raw GCMs data as the input, such that the low resolution input images match reasonably well high resolution outputs, and bias-correction required is left behind [5], [6], [14], [15]. Some works use relatively shallow convolution layers and their downscaling performance may not be better than classic downscaling approaches [7], [12], [15].

Statistical downscaling looks similar with SISR as both aim at getting higher resolution images from lower resolution images if climate variable data are treated as images [6]. However, several differences come up after a closer look.

1) Inputs and outputs in downscaling are clearly from different sources, such as low resolution forecasts from GCM vs historical climate observations [6]. In SISR, the low-resolution input images and high-resolution target images are arguably from a same source, and the high resolution images are often aggregated to from low resolution images as the inputs [13]. Most deep learning based downscaling techniques focused on a single source [5], [15].
2) Statistical downscaling often has additional auxiliary variables. Rainfall events are often associated with other climate variables, e.g., intense low pressure systems and topographical information [6], [7], [12], which are found often beneficial for downscaling [6], [7].
3) Bias and displacement in space or time are common in climate forecasts, especially for precipitation, due to the inherent complexity in climate modelling. Providing multiple possible forecasts is a practice standard for short or long lead time forecasts [1], [3]. Therefore, downscaling performance should be evaluated in terms of both forecast accuracy between two images and forecast skill as ensemble forecasts [**?**], [16]. The later has never been used in deep learning literature as far as we know, while is predominant in climate communities [**?**], [**?**], [16], [17].

To address these differences, we select Very Deep Super-Resolution (VDSR) [18] from latest SISR techniques as a suitable candidate for our downscaling problem based Continuous Ranked Probability Score (CRPS), a widely used ensemble forecast skill metric [**?**], [**?**], [16], [17]. To improve its downscaling performance, we incorporate other resolved climate variables into VDSR and propose Very Deep Statistical Downscaling (VDSD) model. VDSD structure is finalised based on CRPS on a randomly selected validation data set. 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 II and present climate data in Section III. We select three SISR models, and propose and finalise the new downscaling model VDSD in Section IV. Cross-validation results and comparison are given in Section V. We conclude the paper in Section VI with discussions on further developments of VDSD for future operational use.

II. Related works

A. Image Superresolution

Single Image SuperResolution (SISR) is basically the recovery of a low resolution image to a high resolution image. 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 [13]. A lot of superresolution 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.

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 [5].

Since the seminal work by Dong et al. [19], deep learning based SISR techniques have been widely developed and successful applied. Most of them are based on Convolutional Neural Network (CNN) [6], such as Super-Resolution CNN (SRCNN) [19]. As surveyed in [13], these SISR models use a lot of network design techniques, such as residual learning in Very Deep SuperResolution [18], attention mechanism in Residual Channel Attention Network (RCAN) [20], encoder-decoder network or generative adversarial network (GAN), in SuperResolution GAN (SRGAN) [21]and Enhanced SRGAN (ESRGAN) [22], and so on. Because of their superior performance in image superresolution [18], [20], [22], we will examine their downscaling performance based on ensemble forecast skill metric first. In addition, we will incorporate extra inputs into Eq 1.

B. Downscaling techniques for climate forecasts

A large body of downscaling techniques are available in literature, from dynamical downscaling, statistical downscaling, and recent deep learning based downscaling [6].

Dynamical downscaling, via a Regional Climate Model (RCM) forced by boundary conditions from a GCM to run finer resolution simulation, provides a solution to improve spatial resolution. The application of such a 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. Cost-effective 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 [8]. Traditional statistical downscaling techniques, especially for precipitation, have been developed with their own advantages and challenges for different applications, such as Model Output Statistics (MOS), Perfect Prognosis (PP), and weather generators [8]. MOS use statistical approaches to enhance a climate model prediction accuracy by linking the distribution of GCM output to the distribution of observed local-scale climate variables [**?**], [8]. A typical example is Quantile Mapping which assumes that the distribution of model simulated data should preserve the distribution of observed data [**?**], [23]. 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 [8]. QM often has nice performance, and often used in operation [24]. PP is based on empirical relationships established between informative large-scale atmospheric variables (features) and local/regional variables of interest (predictands) such as precipitation [7]. The effective features vary from a location to another, and often required to be extracted in advance of modelling. 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 [12]. These traditional downscaling methods often perform on a single grid point, and maybe time consuming, especially for ensemble SCFs with 10+ members. Comparisons of traditional statistical and dynamical downscaling suggest that neither group of methods is clearly superior, however in practice computationally cheaper statistical methods are widely used [7], [24].

The pre-engineered features and relationships prior to the statistical downscaling modelling process limits their ability to capture important information beyond what we have known [7]. Automatic feature extraction and selection integrated in 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 [9]. Deep learning has successfully been used in precipitation nowcasting [11] and precipitation parameterisations from GCMs [12]. More related to our objective, downscaling techniques are developed based on SISR models [13]. As one of the first works, DeepSD is proposed, which stacked SRCNN together for climate projection downscaling [5]. For long-term climate projection, Rodrigues et al. [14] proposed a very deep CNN-based SISR strategy to interpolate low resolution 125km weather data to 25km output for weather forecasts. [7] assessed CNN methods with three convolutional layers followed by different connection layers for downscaling 200km reanalysis precipitation to 50km observational grids over whole Europe. Wang et al. [15] proposed, based on 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 bias-correction required in downscaling [15]. 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 input for downscaling. YNet was tested on monthly mean of precipitation [6], which has different characteristics of 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 [5], [14], [15]. Some works use relatively shallow convolution layers and their downscaling performance may not be better than classic downscaling approaches [7], [12]. 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.

III. Climate data and pre-processing

A. ACCESS-S1 seasonal 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) [3], [24]. 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 0.6Â°, nearly 60×60km, as the Stochastic Kinetic Energy Backscatter scheme [25] is adopted, which leads to an irreparable grid-scale perturbations [26]. The hindcast[1] data of ACCESS-S are available to the public[2] for the period of 1990-2012 (i.e. 23 years). In each year, it has forecasts on 48 different initialisation dates (i.e. 1st, 9th, 17th, and 25th of every calendar month). Its forecasts have a lead time of 0-216 days, and 11 ensemble members. Each individual 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 9am to 9am (local time). Three precipitation forecast images for 7 Jan 2012 are illustrated in Fig 1.

ACCESS-S1 data also provides a calibrated version. As described in [24], for each forecast initialisation date, each different lead times, and grid point location, it has a calibrated function to downscale to 5km resolution [24]. 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. It is different from the bias-correction spatial disaggregation, which prefers QM and then spatial interpolation [6]. We will use the calibrated data forecast skill assessment and compare it with our downscaling techniques. As its core technique is quantile mapping, and we use QM to indicate it.

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

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

B. BARRA Reanalysis Data

The Bureau of Meteorology Atmospheric high-resolution Regional Reanalysis for Australia (BARRA)is one of regional numerical climate forecast model using the Australian Community Climate and Earth-System Simulator – Regional (ACCESS-R), also Australia's first reanalysis model of the atmosphere [27]. Through assimilating local surface observations and locally derived wind vectors which 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 reaches a good tradeoff between the spatial resolution and consistency with precipitation observations, as recently assessed in [28]. Its spatial resolution of 0.12Â°, around 12km×12km, is realised in the whole region of Australia and New Zealand. 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.

C. Preprocessing

We choose a region from 9Â°S to 43.7425Â°S, from 112.9Â°E to 154.25Â°E as our study region, which covers the whole Australian land (see, e.g., Fig. 1). As pre-processing, we crop all the climate variable surfaces to the same area defined in case study region. These climate variables have different value ranges. For example, precipitation ranges from 0 to 900mm per day, and is about 0-900 mm 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 grid are around 60 km. 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 BARRA 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 saving training time, we only use the first seven lead time forecasts for each initialisation date.

IV. Deep Learning for Ensemble Forecasts

A. Network selection for downscaling ensemble forecasts

GCMs are always substantially simplification of the real world climate system, which is complex and high-dimensional [8]. 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 [**?**], [16]. It is defined as $CRPS\left(\hat{y}_{j,k}^{(i,l,e)}, y_{j,k}^{(d)}\right) = \int_{s=0}^{1}\left(F_{j,k}^{(i.l)}(s) - \mathbb{I}\left(s \le y_{j,k}^{(d)}\right)\right)^2 ds$ 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 spread. 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 generate an aggregated forecasts from an ensemble of forecasts like did in [6], 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 agricultural applications, such as feeding into biophysical models [**?**], [29]. Thus, our downscaling problem is 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 from $\left\{\hat{y}_{j,k}^{(i,l,e)}\right\}_{e=1,\cdots,E}$ and $w_{j,k}^{(i,l)}$ is the weight for the ensemble forecast on Date i, lead time l at location (j, k). We use $w_{j,k}^{(i,l)}$=1 for this study.

To find 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 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 superresolution and employ three SISR models, VDSR [18], RCAN [20], and ESRGAN [22]. They are chosen because of their outstanding performance [13], [22]. The training is based on image superresolution, 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 loss function, calculating error between high resolution images and super resolution image (output from of Eq 1); λ is trade-off parameter and $\Phi(\theta)$ is regularisation term.

ESRGAN's average CRPS on the validation data is not as good as QM, especially for the first 31 lead times. For our downscaling data set, RCAN is hard to converge, 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. Its validation forecast skill 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 tried 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 [18] for image superresolution.

B. Very Deep Statistical Downscaling

As we discussed earlier, other climate variables such as temperature or pressure could influence precipitation and have often been used for precipitation simulation and downscaling [7], [12]. 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 [1], [7],

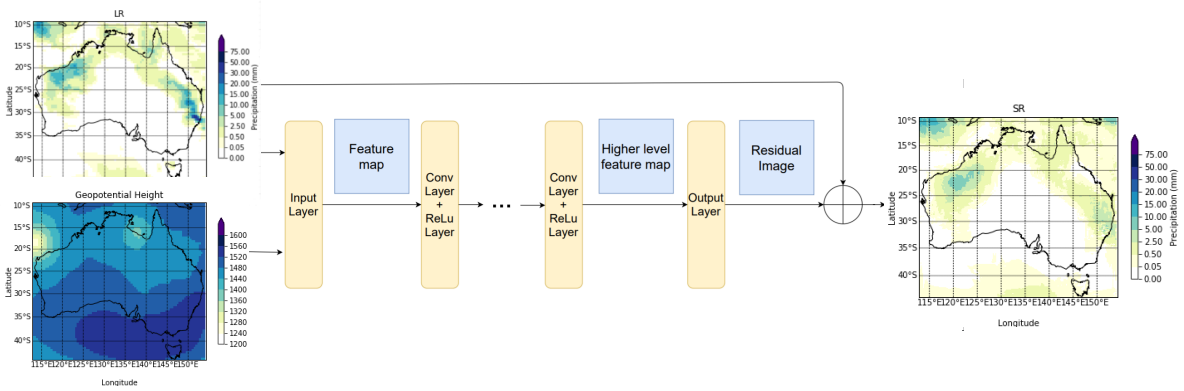

Fig. 2: The structure of VDSD model, modified from VDSR [18], where $\oplus$ represents element-wise matrix addition with input precipitation image, orange blocks are layers of neural network, blue rectangles are feature maps, input and output images are on left and right hand side.

[12]. The overall structure of VDSD is shown in Fig 2, where Geopotential Height (GZ) at 850 hPa is used as the additional input. This GH 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 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, and after multiple intermediate convolution and activation layers, it generates 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 [13], [18].

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 ReLU function to force negative values from the feature maps to be zero. The operation can be formulated as

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

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. 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 of the size of high-resolution image. Suppose the size of input image is $4 \times m \times n$, then the size of the feature map generated by input layer is be $64 \times m \times n$. These basic features then go through multiple intermediate blocks. The intermediate blocks are identical and each of them consists of a convolutional layer, which extract deeper features, and a ReLU layer, which forces negative values to be zero. Each convolutional layer has 64 kernels to produce 64 feature maps and takes 64 feature maps from previous block as input. Therefore, the operation of each intermediate block is the same, which can be written as

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

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 exptected to improve downscaling performance.

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 on the random validation data set to finalise structure of VDSD. We test two different 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. Adding more climate variables than geopotential height improves very little CRPS, and sometimes deteriorate its 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.

V. Cross-validation and Comparison Results

To illustrate the downscaling performance of VDSR and VDSD we have finalised so far, we conduct two leave-one-year-out validation. (1) We took forecasts made on 48 initialisation dates in 2012 for validation and the other forecasts made before 2012 for the training for downscaling methods.

(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. Note that actually more than three years of daily rainfall images are used in cross-validation. The ACCESS-S1 forecasts made on 25th December cover up to 29 July next year for its 216-days lead time forecast.

A. Performance Metrics for Forecasts

A benchmark for long lead time rainfall forecast for a given year is to use observations in 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 [**?**], [**?**]. In this study, we used 1990-2012 as the base period, and thus there are 22 ensemble members in our climatology ensemble forecast.

As we discussed in Section IV-A, 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. To further compare with climatology, we calculate the CRPS skill score for model m against the CRPS of climatology as following.

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

The model with 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 prefect 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 climatology model, and vice versa.

As downscaling techniques based on deep learning are often assessed by MAE [6], [13], we also use another comparison metric, average MAE, which is defined as $\overline{MAE}_{k,j} = \frac{\sum_{f,e}\left|\hat{Y}_{k,j}^{(f,l,e)} - Y_{k,j}^{(d)}\right|}{\sum_{f,e} 1}$ 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

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

Similarly, higher MAE skill score is preferred. These skill score calculation exclude prediction on the ocean.

B. Results for forecasts on 48 initialisation days in 2012

As illustrated in Fig 3 for three ensemble members forecasted on 1 Jan 2012 for 7 Jan 2012, VDSR keeps similar precipitation area shapes 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.

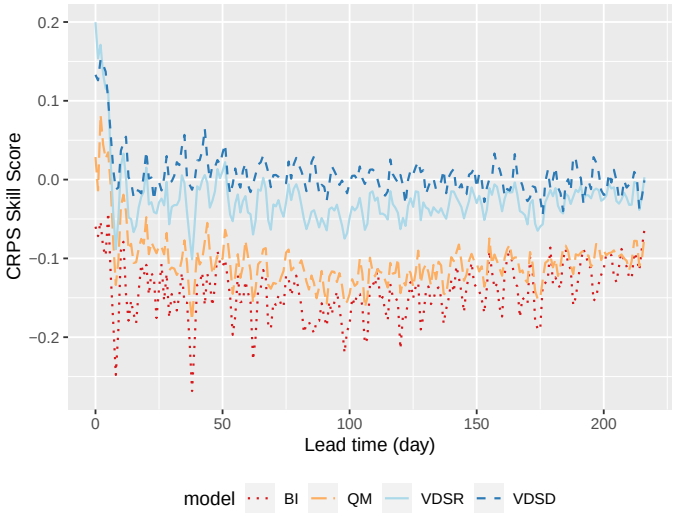

Fig. 3: Average CRPS Skill Scores across Australia for forecasts made in 2012

VDSD could adjust the precipitation area shapes. For ensemble members 2 and 11, VDSD substantially reduces the precipitation along the 30Â°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 IV-A illustrates mean CRPS skill scores across the whole Australia of four downscaling models along with lead time. VDSR has the highest scores in the first 3 lead times, and then VDSD becomes the best of four models for most other lead times. Averaging across the 217 lead times, their mean CRPS skill scores are 5.63×10^{-3},-2.54×10^{-2},-1.05×10^{-1},-1.42×10^{-1}, respectively, for VDSD, VDSR, QM and BI. VDSD is about 3% better than VDSR, and more than 11% better than the both traditional downscaling techniques. Among the four downscaling techniques, only VDSD is better than climatology on average as its positive CRPS skill score. Along with 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 of VDSR and QM is 0.86. The difference is caused by including the extra climate variable geopotential height.

To check the performance of these downscaling techniques for sub-seasonal forecasts, Fig 4 illustrate the average CRPS skill scores for the first 45 lead times (with lead time up to 44 days). The skill scores of BI are around -0.1 for most locations on Australian land. QM has some improvement to be around -0.07. For most locations, VDSR has skill scores near 0. VDSD has positive skill scores for most locations in mainland Australia, and its average is around 0.02. VDSD still has negative skill scores along coastlines in eastern and north-western regions, as well as 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, which indicates the ACCESS-

(a) (b)
BI QM

(c) (d)
VDSR VDSD

Fig. 4: Average CRPS Skill Score for lead time 0 to 45 days across Australia

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

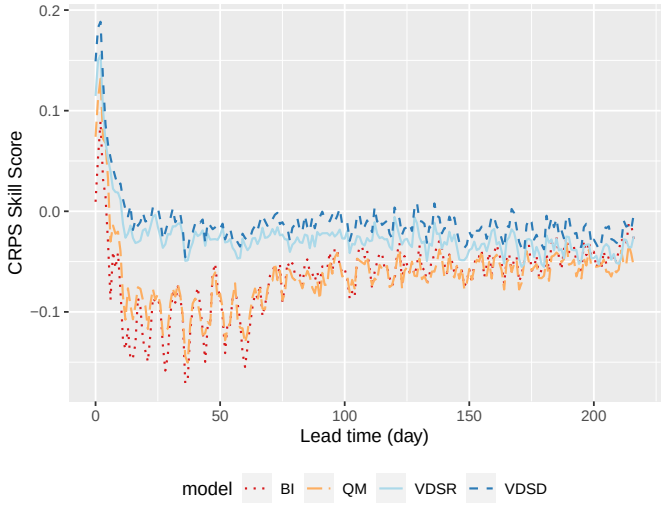

Fig. 6: Mean CRPS skill scores across Australia for forecasts made in 2010.

(a) (b)
BI QM

(c) (d)
VDSR VDSD

Fig. 7: Average CRPS Skill Score for lead time 0 to 45 days across Australia

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

C. Results for forecasts on 48 initialisation days in 2010

Fig 6 illustrates mean CRPS skill scores along with lead time based on 48 seasonal forecasts made in 2010. Along the 217 lead times, VDSD normally has highest CRPS skill score. The mean CRPS skill scores across the 217 lead times are -1.02×10^{-2},-2.53×10^{-2},-6.45×10^{-2}, -6.52×10^{-2}, respectively, for VDSD, VDSR, QM and BI. VDSD is about 0.02 higher skill score than VDSR, and 0.07 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 percent of CRPS skill score [**?**], [17]. For the first 45 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 around 0 skill score though VDSD has slightly higher CRPS skill scores in north Queensland. .

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 2010.

D. 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. Learning rate was 0.0001, and relatively small as the network is very deep, large learning rate may cause vanishing/exploding gradients problem [30]. Optimisation method is Stochastic Gradient Descent (SGD) with 0.9 momentum. Our implementation is written in python (v3.7.4). The implementation is accessible on github (for double blind review, a detailed link is removed).

Table I specifies the hardware we used in the experiments. Training was run on Gadi, a high performance computer in National Computational Infrastructure (NCI), Australia. Forecast validation was done on a normal PC with a GPU.

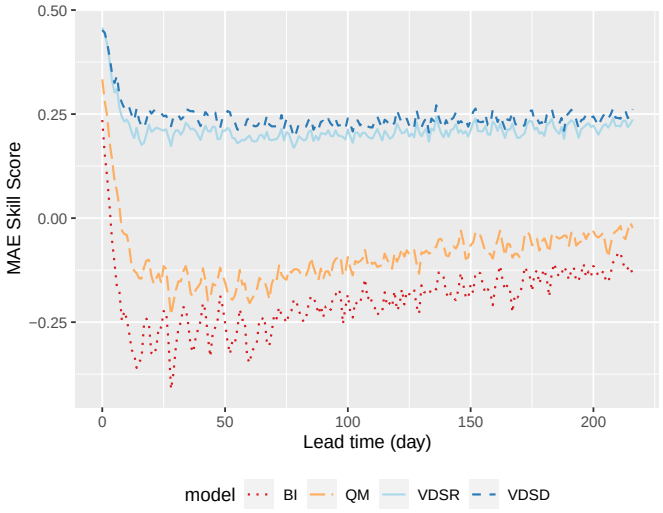

Fig. 8: Mean MAE skill scores for whole Australia for 2010-2011 precipitation forecast.

	HPC(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

TABLE I: Hardware configuration used in the experiments

Table II 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 doesn'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.

VI. Conclusion and Discussions

To improve the downscaling techniques for long lead time ensemble daily precipitation forecasts in Australia, we have applied several representative Single Image SuperResolution (SISR) techniques and selected Very Deep SuperResolution (VDSR) as the suitable deep learning model based on ensemble forecast skill metric, Continuous Ranked Probability Score (CRPS) on a random selected validation data set. We have further incorporated extra climate variables into VDSR, and established Very Deep Statistical Downscaling (VDSD). Both deep learning models have been finalised based on CRPS on the random selected validation data set. On leave-one-year-out cross validation for years 2012 and 2010, VDSD have outperformed VDSR and two traditional downscaling techniques in terms of both forecast accuracy and ensemble forecast skills. 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 could downscale long lead time daily precipitation very fast while needs a lot of time for model development and training. Thus VDSD demonstrates its potential for possible operational use in future.

TABLE II: Computation time 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

For validation results for forecasts made in 2010, the overall average ensemble forecast skill of VDSD is slightly worse than climatology. There are some possible reasons. (1) For forecasts made in 2010, observations from Jan 2010 to July 2011 were 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 well above the long-term average of 465 mm due to the La NiÃ±$aeventpeak$[3].$TheLaNi\tilde{A} \pm aeventpeakin2012ismuchweaker, andmade2012relativelyeasiertofo$
$S1mayperformworsein2010than2012.Forbothyears, VDSDmadesu$
$S1.(3)Theclimatologybenchmarkwehaveusedhave22ensemblemembe$
$1001[17].Therefore, althoughtheCRPSskillscoreisnegativeonaverag$

There are several directions to move the proposed technique for daily operation in future. Station-based precipitation observations have not assimilated in BARRA and its grid precipitation may be not very consistent with on-the-ground observations [28]. To remove such inconsistency, station-specific downscaling techniques like QM may further improve long lead time forecasts. As the spatial and cross-variable relationships may change with time, we will investigate separate downscaling models for different seasons, which is often very helpful in practice. 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 put 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 subject to future work.

References

[1] 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.

[2] The Centre for International Economics, "Analysis of the benefits of improved seasonal climate forecasting for agriculture," tech. rep., Managing Climate Variability Program, 2014. Accessed in Nov 2020.

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

[3] 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.

[4] 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.

[5] 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*, pp. 1663–1672, 2017.

[6] 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*, pp. 3145–3153, 2020.

[7] 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.

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

[9] 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.

[10] 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*, pp. 5617–5627, 2017.

[11] 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.

[12] 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.

[13] 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.

[14] 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)*, pp. 415–422, 2018.

[15] 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.

[16] 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.

[17] 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.

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

[19] 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*, pp. 184–199, Springer, 2014.

[20] 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)*, pp. 286–301, 2018.

[21] 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*, pp. 4681–4690, 2017.

[22] 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*, pp. 63–79, 2018.

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

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

[25] 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.

[26] 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.

[27] 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.

[28] 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.

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

[30] 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.