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

Double Blind Review
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

Double Blind Review
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

Double Blind Review
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—Skillful and high resolution daily weather forecasts for upcoming seasons are of huge value to climate sensitive sectors, especially for agricutlure and construction sectors, to mitigate weather risk and improve production. Global Climate Models (GCM) are providing long lead time ensemble climate forecasts routinely while require downscaling techniques to improve their spatial resolution and consistency with local weather 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 on image superresolution have successfully been developed while most of them focused on simplified situations where low-resolution images can arguably match well high-resolution ones, which are not the case in long lead time daily rainfall forecasts. After applying Very Deep Super Resolution (VDSR) network for long lead time precipitation downscaling, we then propose Very Deep Statistical Downscaling (VDSD) model via including a resolved climate variable geopotential height. An adapted structure of residual learning is used to extract features from geopotential height fields. Tested on real-world downscaling operation, which downscales 60km ACCESS-S1 rainfall forecast up to 217 days lead time to 12km BARRA-R reanalysis data for whole Australia. Leave-one-year-out cross-validation results illustrate their higher forecast accuracy and skill, measured by RMSE and CRPS respectively, than traditional downscaling techniques spatial interpolation and quantile mapping. The results further show that VDSDd perform better or similarly as climatology, a benchmark for seasonal rainfall forecast. In addition, VDSD are lightweight enough to run on modest computer systems, which is favourable for seasonal forecast operation.**

***Index Terms*—Image superresolution, statistical downscaling, seasonal climate forecasts, ensemble forecast**

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 construciton [1]. For example, multiple month lead time daily rainfall forecasts can benefit the whole agriculure value chains, such as helping farmers adapt their farmming planning and management, and insurers and traders adjust their pricing scheme. Skilful SCF are estimated to contribute around $60 per hectar per year for cotton, $40 for wheat or sugar cane in Australia [2]. For the whole Australia, potential annual value added from skillful SCF would be around $1.6 billion for the agricultural sector and $192 million for the construction sector [3]. To fulfill these potential values, SCFs should be provided timely and skillful 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], [4]. They couple physics-based models of the ocean, atmosphere, land surface and sea-ice, and could 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-constrainted. Limited by computational resources, coarse spatial resolution and simplified nature of GCMs often lead them to produce data not reliably consistent with observed weather, especially for preciation. To improve forecast skills and quantifying uncertainty, ensemble forecasts, i.e., multiple simulations of a single model each with slightly different conditions, are normally carried out and published. For example, the operational SCFs from Australia's Burean of Meterology (BoM) have 11 ensemble members for each forecast issue date [4] and from United States' National Centres for Enironmental Prediction have 40 members [7]. The coarse spatial resolution and limited skills in representing local climate characteristics of GCMs circumvent the direct use of outputs in above-mentioned weather-sensitive sectors applications [8]. The barriers are normally bridged via downscaling techniques which generates more skilful 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 dyanmic downscaling, statistical downscaling, and recent development on deep learning-based downscaling [9]. Dynamic downscaling, via a climate model forced by boundary conditions from a GCM to run finer resolution simulation, provides a solution to improve spatial

The work was funded by XXX

resolution. The application of such regional climate modelling is computational expensive and downscaling performance depends on the domain of interest, especially for high-resolution (with grid spacing ˜10 km or less) simulations over large domains [10]. 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 [9]. 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 [9].

MOS use statistical approaches to enhance a climate model prediction accuracy by linking the distribution of physical climate model output to the distribution of observed local-scale climate variables [9], [11]. A typical example is Quantile Mapping which assumes that the distribution of model simulated data should preserve the distribution of observed data [11], [12]. PP is based on empirical relationships established between informative large-scale atmospheric variables (features) and local/regional variables of interest (predictands) such as precipitation [8]. The effective features vary from a location to another, and often required to be extracted in advance of modelling. The feature extraction procedure could capture little useful information beyond our prior knowledge, and is very time consuming as GCM's outputs are high-dimensional [13]. PP modellling techniques build mainly on (generalized) linear regression and analog techniques [8]. Weather generators use stochastic models to generate Monte-Carlo time series simulations resembling the observed statistical characteristics, which explicitly model temporal dependence [14]. 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 [8].

The inadequacy of the traditional statistical/machine learning downscaling techniques may stem from the pre-engineered features and relationships prior to the modeling process, rarely exploiting their spatio-temporal dependencies exhaustively. This limits their ability to capture important information beyond prior knowledge [8]. Automatic feature extraction and selection integrated in the modeling process with convolutional neural networks (CNNs) has achieved notable success in modeling data with spatial context, recently on climate science [15]. CNNs have found applications in precipitation nowcasting [16], [17] and precipitation parameterisation from GCMs [13]. More related to our objective, downscaling techniques are developed based on super-resolution convolutional neural networks. As one of the first works, [5] proposed DeepSD, which stacked SRCNN together for climate projection downscaling. For long-term climate projection, [18] proposed a very deep CNN-based super-resolution strategy to interpolate low-resolution 1.25 degree weather data to 0.25 degree output for weather forecasts. [8] evaluated CNN methods with three convolutional layers followed by different connection layers for downscaling 200km reanalysis precipitation to 50km observational grids over whole Europe. [19], based on deep convolutional neural network with residual blocks and batch normalizations, proposed Super Resolution Deep Residual Network (SRDRN) for downscaling daily precipitation and temperature. [6] presented YNet, consisting 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 data. It was tested on monthly mean of precipitation, which has different characteristics of daily precipitation. These pioneering deep learning based downscaling works have various success. To reuse techniques developed in image superresolution based on deep learning, some works oversimplified the downscaling problem by taking the upscaled observation rather than GCMs data as the input, such that the low-resolution inputs match well with high-resolution outputs, and leave bias correction behind [5], [18], [19]. Some works use relatively shallow convolution layers and their downscaling performance may not be better than classic downscaling approaches [8], [13].

Statistical downscaling is very similar to image super-resolution (SR) as both aim at getting higher resolution images from lower resolution images if we regard climate variable data as images [6]. But statistical downscaling are different from three aspects.

1) Input and output are normally from different sources, such as low-resolution forecasts from GCM and historical climate observations/simulations [6]. In image super-resolution, the low-resolution input images and high-resolution target images are arguably from the same sources, normally the high-resolution images are aggregated to low-resolution images as the inputs for image superresolution.
2) Statistical downscaling often has additional auxiliary variables. Rainfall events are often associated with other climates, e.g., intense low pressure systems, topographical information, and climatological information [6], [8], [13], which are often informative for downscaling [6], [8].
3) Bias and displacement in space or time very common in climate forecasts, especially precipitation, due to the problem inherenat complexity. Providing multiple possible forecasts, instead of one, is a operational standard for short or long lead time weather forecasts [1], [4]. Thus, downscaling performance should be evaluated in terms of accuracy between two images and the whole ensemble (see e.g. in Fig 2) [20], [21]. The later has rarely been done in deep learning literature while predominent in the climate and hydrology communities [9], [20], [21].

Precipitation raw forecasts from climate models are partially parameterised and are usually considered less reliable compared to the directly resolved variables, such as pressure and temperature [13]. To improve downscaling performance, we

will adopt an image superresolution (Very Deep SuperResolution (VDSR)) technique to take into an extra input, daily 850hPa geopotential height, called it Very Deep Statistical Downscaling (VDSD). VDSD is tested on real-world application scenarios and evaluated both downscaling accuracy and ensemble forecast skills. Cross-validation results illustrate its better performance than classical downscaling technique as well as direct application of VDSR. In addition, VDSD outperforms climatology, a benchmark for long lead time climate forecasts.

[9]

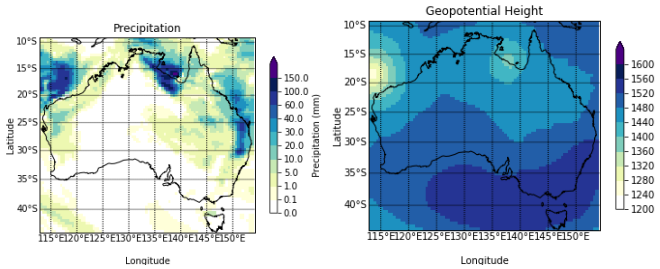


Fig. 1: Demo of two channels of low-resolution input image, daily forecast of 25/01/2012

II. Data Sets used and pre-processing

A. ACCESS-S1 seasonal forecast data

We use daily rainfall forecasts from Australia's operational seasonal climate forecast system, the Australian Community Climate and Earth-System Simulator (ACCESS) Seasonal model version 1 (ACCESS-S1) [4]. It is the Bureau of Meteorology's climate modelling system used for outlooks on multi-week through to seasonal timescales. Its development is based on the United Kingdom Met Office's Global Seasonal forecast system version 5 model configuration 2 (GloSea5-GC2). ACCESS-S1 couples the state-of-the-art land surface model, ocean model and atmosphere model. Its atmosphere model has enhancements to the ensemble generation strategy to make it appropriate for sub-seasonal forecasting, and a larger ensemble size. The resolution of the atmospheric model is raised to 0.6 degree, nearly 60km*60km, as the Stochastic Kinetic Energy Backscatter scheme [22] is adopted, which leads to an irreparable grid-scale perturbations [23]. The hindcasts of ACCESS-S rainfall are available for the period of 1990-2012 (i.e. 23 years) with a lead time of 0-216 days, 48 different initialisation dates (i.e. 1st, 9th, 17th, and 25th of every calendar month) and 11 ensemble members. Each individual forecast member provides a full description of the evoluation of weather for upcoming 217 days, and collectively these ensemble forecasts indicate the likelihood of a range of future weather scenarios. [4] showed that ACCESS-S1 obtains overall skill improvement on rainfall, maximum temperature (Tmax) and minimum temperature (Tmin) over Australia on multi-week scales compared to its ancestor POAMA.

B. BARRA-R 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 ACCESS-R, also Australia's first reanalysis model of the atmosphere [24]. Using ACCESS and 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. BARRA-Rcovers all of Australia, New Zealand and the maritime continent reaches a good tradeoff between the spatial resolution and consistency with precipitation observations [25]. The spatial resolution of 0.12 degree, around 12km * 12km, is realised in the whole region of Australia and New Zealand. Su. et al. think the traditional downscaling method such as statistical downscaling and dynamic downscaling usually involves embedding a new weather model on top of a global climate model, ESM or CSM, which generates high-resolution predictions. They also think the reanalysis method is an extension and evolution of these two traditional downscaling methods. And unlike both of them, the way of observations that are used in regional reanalysis and in global high-resolution simulations are the same. They both minimise model errors. Frequency distributions, extreme values, and actual space-dependent and time-dependent variability can be well represented by reanalysis. BARRA use The Unified Model [26], 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. [25]. Six-hour accumulated precipitation, obtained from BARRA from 1 January 1990 to 31 December 2013, is aggregated to daily frequency by taking the sum of the four 6-h gridpoint values within each 24-h window. All of the daily aggregation is based on the SILO climate day definition of 9am to 9am (local time).

C. Preprocessing: Cropping and Normalization

We chose the region from 9°S to 43.7425°S, from 112.9°E to 154.25°E as our case study region, which covers the whole Australian land (Fig. 2)

Crop all the climate variable surfaces to the same area defined in case study region. The raw model forecast data on the native atmospheric grid, around 60 km, are first interpolated to the higher-resolution grid using bilinear interpolation.

The SRDRN method requires a data preprocessing step, including data normalization and precipitation data augmentation. Specifically, each variable is normalized by subtracting the mean μ and dividing by the standard deviation σ. Here μ and σ are scalar values that are calculated based on the flattened variable for the entire data set.

In our climate forecast downscaling model, the first channel represents the predicted amount of rainfall and the second is the geopotential height. These two channels have significantly different range and average values. The range of precipitation is about 0-900 mm per day, while the range of geopotential height is around 1200-1600. To bring climate variables to have

similar value ranges, we do normalisation at each input image set, such as, for geopotential height

$$gh^* = max(PR) * \frac{gh - min(gh)}{max(gh) - min(gh)} \tag{1}$$

where gh* and gh represents the new value and original value of each pixel in geopotential height channel, PR and GH is the whole precipitation channel, and original geopotential height channel.

D. Image Superresolution via Interpolation

Super resolution is basically the recovery of low resolution images to high resolution images. low-resolution image LR is regarded as the result of degradation:

$$LR = \mathcal{D}(HR; \gamma) \tag{2}$$

,where D is degradation mapping function; HR is high-resolution image corresponding to LR; γ is the parameters of the degradation mapping function. A lot of super-resolution data sets are obtained by Degradation Mapping. In this way, a series of low - and high-resolution images have been created, and a lot of scientists are working on the following process, (i.e. super-resolution) :

$$SR = \mathcal{F}(LR; \theta) \tag{3}$$

where $\mathcal{F}$ is the super-resolution mapping function and θ is the parameters of $\mathcal{F}$. From above, all the work is to find the function $\mathcal{F}$ in the equation 3.

1) Interpolation-based Approaches: In real life, we often encounter in the geometry space can be enlarged image, transformation and so on, the condition of these operations are needed in the source image and target image to establish a mapping rule between

$$(x', y') = T(x, y)$$

, makes the two set up a corresponding relationship between image pixel coordinates, for each pixel of the image target assignment.

The mapping from the source image to the target image is called forward mapping, but this mapping method may cause two problems: several pixels of the source image is mapped to the same position in the target image; Some locations in the target image do not have pixels to assign. This leads to the problem of how to convert multiple output values into a single output value and how to assign values to pixels that are not mapped to the target image. And we often use to map, this method is more efficient than the forward mapping, it is from target image to the mapping of the source image, namely $(x, y) = T^{-1}(x', y')$, thus avoiding the problems that exist in the forward mapping, but also inevitably exists some back to the source image pixel mapping problem of coordinates of floating Numbers, and the image pixel values only at integer coordinates are defined, At this point, the image interpolation method is used to assign a value to the pixel in question. Common interpolation algorithms include nearest-neighbour interpolation, bilinear interpolation, bicubic interpolation and so on.

a) Bilinear Interpolation: The bilinear interpolation method is improved from the nearest interpolation method, but the difference between them is not big. The nearest interpolation method only takes the value of one pixel, while bilinear interpolation, using weighted average algorithm and linear equation:

$$y = y_1 \frac{(x - x_2)}{x_1 - x_2} + y_2 \frac{x - x_1}{x_2 - x_1}$$

, where (x, y) is target pixel that need to be solved, $(x_1, y_1), (x_2, y_2)$ is the given value,to calculate the target pixel. Fig shows the basic principle.

b) Bicubic Interpolation: Compared with the first two interpolation algorithms, bicubic Interpolation [27] is the most complex interpolation algorithm, but the image processing performance is better, and become the most widely used interpolation algorithm. In this branch of mathematics, bicubic interpolation is the most common interpolation method in two-dimensional space. Through the weighted average of the nearest sixteen sampling points, The value of the function f at the point (x,y) can be obtained. The interpolation function is cubic polynomial, approximated from function $\frac{sin(\pi * x)}{x}$, which looks like FigII-D1a. And it define as h(x):

$$h(x) = \begin{cases} (a+2)|x|^3 - (a+3)|x|^2 + 1 & , 0 < |x| \leq 1 \\ a|x|^3 - 5a|x|^2 + 8a|x| - 4a & , 1 < |x| < 2 \\ 0 & otherwise \end{cases} \tag{4}$$

where x is the difference between the target pixel position coordinates and four different horizontal and vertical coordinates of 16 points (like Equation 6 and 7), and a usually set to 0.5. Using the interpolation function, we can obtain how much the 16 pixels around the target contribute to the target pixel. then the target pixel can be defined as:

$$f_h(i + u, j + v) = A^* B^* C \tag{5}$$

$$A = [h(u+1), h(u), h(u-1), h(u-2)] \tag{6}$$

$$B = \begin{bmatrix} f(i-1, j-1), f(i-1, j), f(i-1, j+1), f(i-1, j+2) \\ f(i, j-1), f(i, j), f(i, j+1), f(i, j+2) \\ f(i+1, j-1), f(i+1, j), f(i+1, j+1), f(i+1, j+2) \\ f(i+2, j-1), f(i+2, j), f(i+2, j+1), f(i+2, j+2) \end{bmatrix} \tag{7}$$

$$C = [h(v+1), h(v), h(v-1), h(v-2)]^T \tag{8}$$

Overall Equation 6 is equivalent to:

$$f(i+v, j+u) = \sum_{row=-1}^{2} \sum_{col=-1}^{2} f(i+row, j+col) h(row-v) h(col-u) \tag{9}$$

It does not consider a single-pixel around the interpolation point but considers more pixels around the interpolation point.

Therefore, with the help of bicubic interpolation, more accurate pixel values of the interpolation point can be obtained. Compared with the first two interpolation methods, the bicubic interpolation method can reconstruct high-quality graphs. From all the above, the Bicubic Interpolation method becomes one of my baseline models. For easier expression in the next chapter, we use $BI(*)$ to denote the bicubic interpolation operator.

E. Quantile Mapping (QM)

Quantile Mapping (QM) [28] is another significant method which has great performance when the leading time is not large. It learned the matching of an empirical distribution from an 11-day slide window.

Quantile-mapping (QM) is a revised method based on frequency distribution, and it is believed that the observed and simulated precipitation are consistent in frequency distribution [29]. This method corrects the simulated precipitation by establishing the transfer function between simulated precipitation and observed precipitation. The specific methods are as follows

To predict a high-resolution weather forecast on a date, the climatology model would study what happened on the same date in previous years and analyse its relationship using knowledge of climatology [21]. Basically we use these historical observations to form an ensemble forecast. Such forecasts often has nice performance, and performs as a benchmark.

III. Deep Learning for Downscaling Ensemble Daily Rainfall Forecasts

A. Problem of downscaling ensemble forecasts

For low resolution output impages from GCMs, $X^{(e,d,l)} \in \mathcal{R}^{m_0 \times n_0}$ and $Z^{(e,d,l)} \in \mathcal{R}^{m_0 \times n_0 \times p}$ with respect to target high resolution images $Y^{(d)} \in \mathcal{R}^{m \times n}$,we would like to find a function F, which generate the same resolution as $Y^{(d)}$, $\hat{Y}^{(e,d,l)} = F\left(X^{(e,d,l)}, Z^{(e,d,l)}, \theta\right)$, such that minimise $\underset{d,l}{CRPS}\left(\hat{Y}^{(e,d,l)}, Y^{(d)}\right)$.

d indicates the target date; e indicates ensemble member ; l indicates the forecast lead time (0 to 126);

Some training data for (d,l); some validation data to finalise a model;

Choices: layers of network (15 to 25); objective functions L1 or L2

==========Need to discuss with Weifan and Minzhe =========

With the development of computer science, artificial neural network is becoming more and more mature.Starting with the task of identifying regressions, and gradually spreading out, this includes super resolution. The super-resolution algorithm based on deep learning is roughly divided into three types by [30] :supervised SR, unsupervised SR, and domain-specific SR. We know that deep learning network has a learning process.

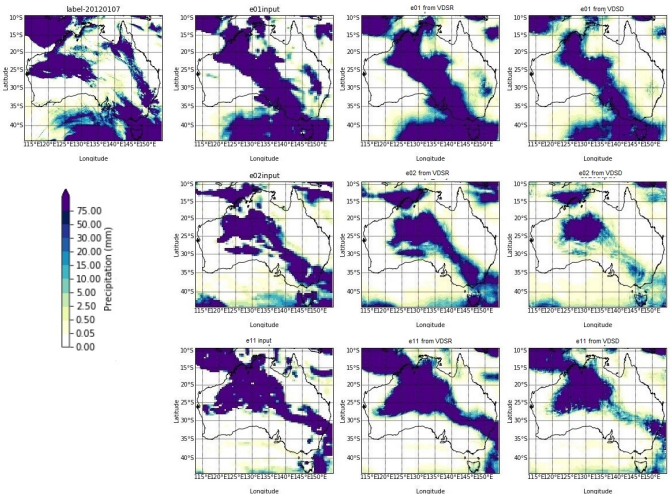


Fig. 2: Demonstration for ensemble rainfall forecasts on 2012-01-07 with lead time one day. Images from the first to the last column are the high-resolution precipitation, ensemble forecasts from a climate model (say, ACCESS-S1), downscaling results of VDSR, and results of VDSD.

Combined with the Equation 3, deep Learning-based problem can be define as:

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}\left(SR, HR\right) + \lambda\Phi(\theta) \tag{10}$$

, L is loss function, calculating error between high-resolution(Ground truth) image and super-resolution image(output of super-resolution mapping function); λ is trade-off parameter and $\Phi(\theta)$ is regularisation term.

for each date, as shown in figure2, all ensemble members share the same label image, which means 11 input image will generate 11 high resolution output images, and the 11 output images will be compared to the same label.

B. Learning relationship via VDSR

For example, a very deep super resolution neural network used residual learning and large learning rate to speed up learning and very deep network to take a large image context into account such that may correct displacement in raw climate forecasts. improved the image quality of the output images [31].

RCAN: failed, partially because of L2 objective used. RCAN failed: very deep very hard to train, patches (hard to catch big picture), edge, attention mechanism,

comprised of convolutional layers, residual blocks, and upsampling blocks (see Fig 1), with data preprocessing a prerequisite step

Typical result of three ensembles are shown in figure 2. These three inputs has the same target date 2012-1-7, and their leading time are all 3, thereby they share the same label.

C. Learning relationship via VDSD

The overall structure of multiple-input VDSR downscaling model is shown in feature 3. This model takes two upsampled

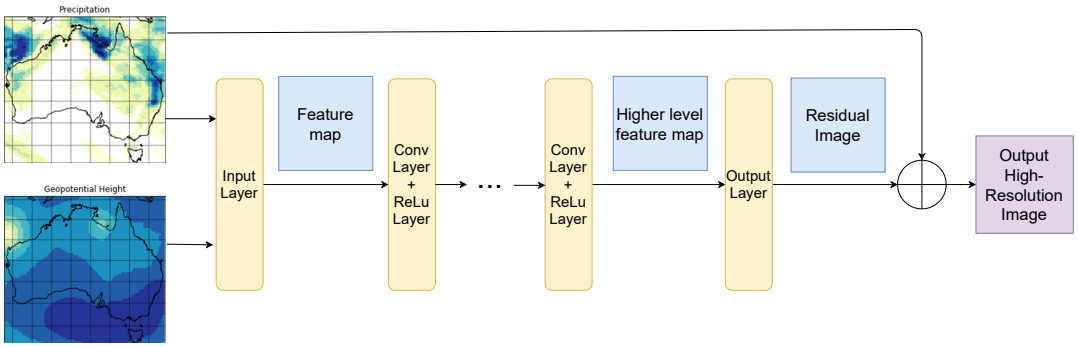


Fig. 3: The structure of VDSD model, where $\oplus$ represents element-wise matrix addition, orange blocks are layers of neural network, blue rectangles are intermediate images, input and output images are on left and right hand side.

low resolution images as input, which includes precipitation and geopotential height daily forecast. For each 2-channel input, it predicts a residual image, which is the difference between low-resolution rainfall forecast and label images, and finally perform element-wise addition to add the upsampled low resolution precipitation image onto the residual image to generate a predicted super resolution image. Two channels are used for training and feature extraction, but only precipitation forecast is used for final addition. In this way, the model does not need to predict the whole rainfall forecast but predict the difference between low-resolution forecast and the ground truth value. The high-level formulation can be written as

$$SR = PR_{lr} + F(PR_{lr}, ZG_{lr}) \tag{11}$$

where SR is output high resolution precipitation forecast, F represent all operations in the whole neural network, PR_{lr} and ZG_{lr} are upsampled low-resolution forecast of precipitation and geopotential height.

the proposed model has a new adopted version of residual learning. The traditional residual learning model adds residual to the input image to get the result. However, the input for my rainfall forecast downscaling model contain both rainfall and geopotential height images, thereby it would be meaningless to add both input images to the residual. Instead, although both channels are used to extract features, only precipitation input image is added to the residual image. In this way, residual learning can improve the performance, and features from geopotential height data can improve precipitation prediction.

As shown in feature 3, the model mainly has three part: input layer, intermediate feature extraction layer and output layer. The input layer takes two-channel input image extract features with 64 filters. Then the intermediate layers generate higher level feature maps. Finally, the output layer will convert the feature map into a residual image, which is the predicted difference between input and target image. The residual image will be added to the input precipitation image to output a super-resolution rainfall forecast. All procedures and intermediate outputs are shown in figure **??**. Next, these three parts will be introduced one by one and the methods of feature extraction and image reconstruction will be discussed.

The first image is input image, then 64 feature maps are generated, and then Convolutional and ReLU layers extract 64 higher level feature maps, and finally the residual image is constructed, which is added to the input precipitation forecast to get a super resolution forecast

1) Input Layer: There is significant difference between original SISR on general images and its application in climate forecast downscaling. This is because the aim of downscaling is probabilistic ensemble forecast, thereby the model should not only improve the image quality by enhancing the resolution but need to use all ensemble members to predict high-resolution rainfall images that are closed to the target image. The result is not only evaluated by Peak signal-to-noise ratio (PSNR), but by CRPS and CRPS skill score, which can show a possibility that the ensemble members are reliable.

Because there is difference between the forecast and the target image, the model should learn how to improve the forecast while enhancing the resolution, thereby other weather features might be helpful. For example, geopotential height is found to be useful in rainfall downscaling [13]. Geopotential height is similar to geometric height but adjusted with gravity. Given an elevation h, geopotential is defined as

$$\Phi(h) = \int_0^h g(\phi, z)dz \tag{12}$$

where $g(\phi, z)$ is gravitational acceleration under altitude ϕ and geometric elevation z [32]. And geopotential height is defined as

$$Z_g = \frac{\Phi(h)}{g_0} \tag{13}$$

where g_0 is the standard gravity at mean sea level. Geopotential height thereby shows the variation of gravity with the altitude and latitude.

A study in the united states used geopotential height from three levels (500, 850, 1000 hPa) as input to improve the daily precipitation prediction with convolutional neural network [13]. Four channels of climate data were stacked as an

input image to go through convolutional and ReLU layer, and then performed max pooling, followed by a dense layer to reconstruct the high-resolution image. Their result indicates that geopotential height can contribute to rainfall downscaling, thereby using geopotential height may also improve the performance of deep learning based Australian seasonal forecast downscaling model.

As shown in figure 3, the 2-channel input image will firstly go through an input layer. This layer has a convolutional layer and a ReLU layer. The convolutional layer has 64 kernels, and will produce 64 first level feature maps. Then the ReLU layer will perform ReLU function to remove the meaningless negative values from the feature maps. The operation can be formulated as

$$FM_0 = F(PR_{lr}, ZG_{lr}) = ReLU(Conv(PR_{lr}, ZG_{lr})) \tag{14}$$

where FM_0 is the first level feature maps generated by input layer, and ReLU() and Conv() are ReLU and convolutional 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 (2*m*n), then the size of the feature map generated by input layer will be (64*m*n). (In this project, the size of training image is 316*376.) It will convert the two-channel image into a low level feature map as the second image shown in figure **??**. Most of the values in the feature maps are 0 after ReLU layer, and some maps are similar.

This input layer can extract basic features from the 2-channel input and generate 64 feature maps. It also allows the following intermediate layers to extract higher level features.

2) Intermediate Blocks: There are totally 18 intermediate blocks between the input and output layers. All of them are identical and each of them consists of a convolutional layer, which extract deeper features, and a ReLU layer, which removes the negative values. 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

$$FM_n = B(FM_{n-1}) = ReLU(Conv(FM_{n-1})) \tag{15}$$

where FM_n represents the nth level feature map, B is the operation of intermediate block. Therefore, the result of operation of all intermediate blocks can be written as

$$FM_{18} = B^{18}(FM_0) \tag{16}$$

where

$$B^n(x) = \begin{cases} B(B^{n-1}(x)) & , \ n > 1 \\ B(FM_0) & , \ n = 1 \end{cases} \tag{17}$$

and FM_0 represents the first level feature map generated by the input layer.

These intermediate blocks can extract higher level features. The large numbers of layers and kernels allow the model to learn complex patterns. These complex patterns may improve the performance of SISR and precipitation forecast. An eighteenth level feature map is shown in the figure**??**, and this will be used as input of output layer to generate the residual image.

3) Output Layer: Output layer is a convolutional layer that converts 64 high level feature maps into a residual image as shown in figure **??**. That is to use discovered complex patterns to predict the difference between unsampled low-resolution forecast and the target image. Finally, the residual image will be added to the upsampled precipitation input image to predict a super resolution precipitation forecast.

D. Loss Function and Training

The hyper-parameters used in the training are as follows. The number of epochs is 50, which is approximately the time of convergence. Learning rate is 0.0001, and relatively small as the network is very deep, large learning rate may cause vanishing/exploding gradients problem [33]. Optimization method is Stochastic Gradient Descent (SGD) with 0.9 momentum. Loss function is L1 which is defined as

$$L_1 = mean(||SR - HR||) \tag{18}$$

where SR is the predicted super resolution image and HR is the ground truth high-resolution image. The L1 measures the mean absolute difference between values of all pixels. Mean squared error is used in original VDSR setting [31]. Our initial experiments found that for downscaling L1 is more suitable than L2.

Different from [34] where auxiliary inputs have high resolution won't go through feature extraction, geopotential height will generate features via convolution during the training.

E. Verification Metrics

Forecast performance is assessed in terms of both accuracy and ensemble forecast skill. For SISR model, the output super resolution image should be as similar as possible to the label image. Its performance is good when the difference between the pair of images is small. We use Mean Absolute Error (MAE) as metrics

$$MAE = \frac{1}{mn} \sum_{i}^{m} \sum_{j}^{n} |R^{(e)}(i,j) - \widehat{R}(i,j)| \tag{19}$$

R and $\hat{R}$ are observations and downscaling output respectively. Taking climatology as the reference forecast, we can define MAE skill score of each pixel is computed by

$$MAE\_SS = 1 - \frac{MAE}{MAE_{clim}} \tag{20}$$

a) Ensemble Forecast Accuracy: The accuracy of ensemble forecast can be evaluated by the continuous ranked probability score (CRPS) [20]. Each grid point can have a CRPS score on each initialization date and each leading time,

and the mean CRPS score of forecast map of a date t can be formulated as

$$CRPS(t) = \frac{1}{mn} \sum_{m} \sum_{n} \int_{-\infty}^{\infty} \left[F\left(SR_{t,m,n}\right) - \mathbf{1}\left(SR^{g}_{t,m,n} \geq H \right. \right. \tag{21}$$

where m and n is the size of image, t contains leading time and initialization time. To evaluate the CRPS, the mean value over all date in the year is used, so a means CRPS score is computed for each leading time.

Climatology model is used as the reference forecast. The CRPS skill score of each pixel is computed by

$$CRPS\_SS = 1 - \frac{CRPS}{CRPS_{clim}} \tag{22}$$

where CRPS and $CRPS_{clim}$ are the element wise CRPS scores of my deep learning based downscaling model and climatology model, and all CRPS scores have a corresponding leading time.
The model with higher CRPS score tends to have better accuracy. If the skill score is positive on average, then the new model is better than climitology model. If the reference model is the same, the model with higher mean skill score tends to be better.

IV. Cross-validation and Comparison

Calibration is based on the model forecast characteristics obtained from a large set of hindcasts. The calibration is a function of start-date, lead time and location [35].

1) Three periods are chosen to evaluation downscaling performance, 2012-2013, which is the last year of ACCESS-S hindcast data; 2010 is a typical La La Niña which means increased rainfall across much of Australia; 1997 a typical El NiÃ±o year
2) bilinear interpolation (e.g., bicubic) is considered the baseline for spatial downscaling of precipitation fields

All the code was running on the newest super computer called Gadi from National Computational Infrastructure (NCI). The computation resource used for this project is list in the table

The CRPS skill score excludes prediction on the ocean as other downscaling techniques such as QM requires historical observations which may not be available.

The confidence band graph shows the confidence of different downscaling models in comparison to the reference climatology model. As shown in the figure 4, for validation on 2012, the overall skill of the proposed model is slightly better than climatology model, especially when leading time is small. On the other hand, other traditional postprocess methods like interpolation and quantile mapping are worse than the reference, they can have positive skill scores only in first several days. The median value lines from these models are compared in figure **??**. Deep learning base models have better skill than traditional methods. The single input VDSD model has greater skill for small lead time forecasts, but the proposed model overall has a better skill, especially when the leading time is greater than 7. Moreover, my downscaling model

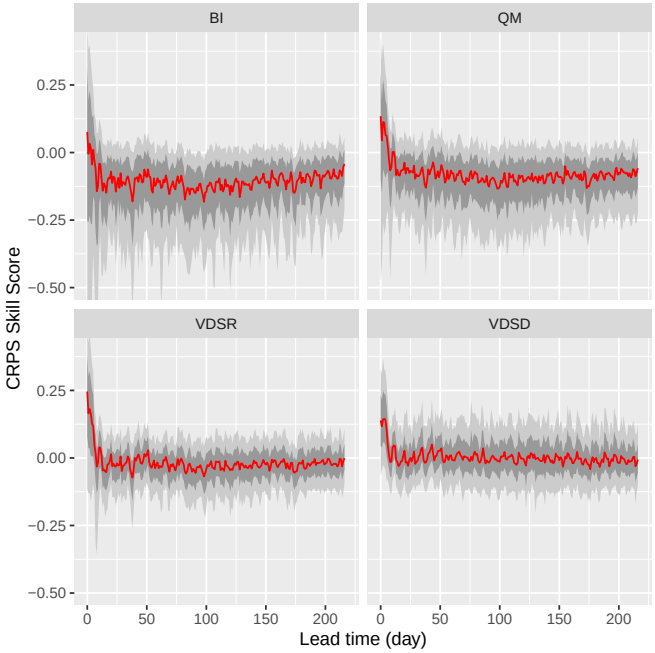


Fig. 4: CRPS skill score distribution of BI, QM, VDSR and VDSD models in 2012 precipitation forecast for whole Australia for lead time 0 to 216 days. The light grey shade indicates the 90% confidence band of CRPS skill scores for all the grid points, the dark grey shade indicates the 50% confidence band, and the red line indicates the median.

Fig. 5: Mean values of CRPS Skill Scores across Australia of four models for 2012-2013

is more stable in predicting 2012 precipitation, because the size of shades, which represent the 50% and 90% confidence intervals are smaller than that of other models.

A. Results for 48 hindcast initialisation days in 2012

The visualization of average CRPS and skill score values are shown in figure **??**. These maps demonstrate the down-

Fig. 6: Mean values of MAE Skill Scores of four models for whole Australia for 2012-2013 precipitation forecast

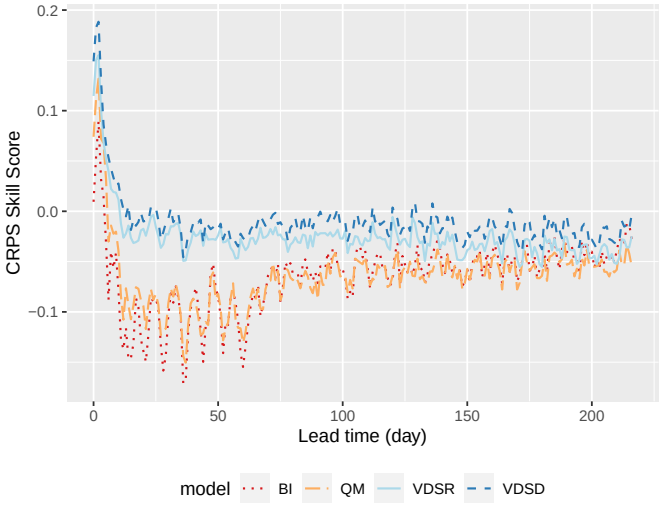
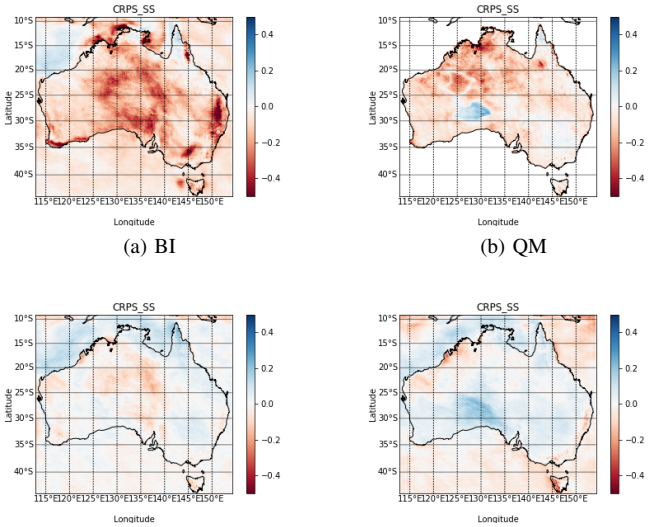


Fig. 8: Mean CRPS Skill Scores across Australia of four model for 2010-2011 precipitation hindcasts



(a) BI (b) QM

(c) VDSR (d) VDSD

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

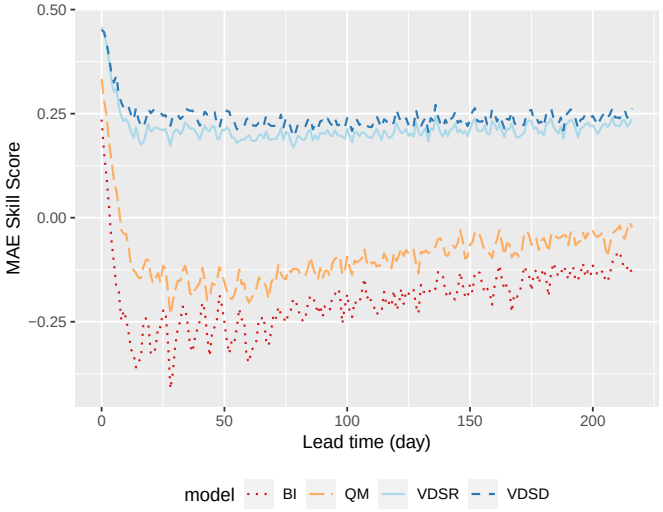


Fig. 9: Mean MAE Skill Scores of four models for whole Australia for 2010-2011 precipitation forecast

scaling performance of different locations. The rainfall on the surrounding ocean is extremely hard to predict for postprocess models, thereby the CRPS is high and skill score is negative on average. The CRPS score is high on the east and north coast. The possible reason is that the amount of rainfall on ocean is large and hard to predict. For CRPS skill score, the proposed model outperforms climatology in southern and northeast parts of Australia but has worse skill on the northwest and Tasmania. The performance of predicting southeastern area is similar. Overall, VDSD prediction model is more accurate than the reference in mainland Australia.

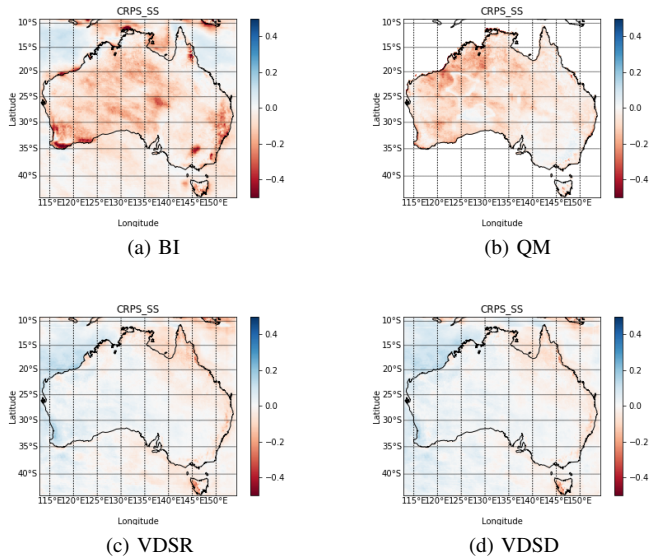
B. Results for 48 hindcast initialisation days in 2010

The CRPS skill score is shown in figure **??**.

The overall skill of VDSD is slightly worse than climatology model. The first reason is that these two years are strong El , thereby the rainfall is extreme and hard to predict compared to 2012. The second reason is that climatology method totally uses 21 ensemble members, and large ensemble size will lead to better CRPS score. Therefore, although the CRPS skill score is negative on average, the proposed model is still promising because it uses less ensemble members and time.

In comparison to other downscaling models like VDSR, the proposed model still has great advantages overall.

The difference of CRPS score is very small, especially



(a) BI (b) QM

(c) VDSR (d) VDSD

Fig. 10: Average CRPS Skill Score for lead time 0 to 45 days across Australia

for 2010 and 1997 prediction. T-test is used to determine whether there is significant difference between model performance. I performed one-side paired t-test by assuming that my model would be better than previous VDSRd model. For 2012 forecast, the p-value of the one-side paired t-test of 76 pairs of CRPS values is 0.017, and the p-value of for 2010 prediction is smaller than 0.0001, which indicates that my model generates statistically significantly better result than VDSR at the significant level 0.05.

The quality of super resolution output image can be visualized and evaluated by PSNR.

Compared to the input and single input downscaling model, my model can adjust the rainfall prediction and make it closer to the label image. For example, there are little precipitation near the 30-degree latitude line, and the proposed model predicted it successfully and reduced the amount of rainfall at that area for all ensembles.

The drawbacks of the proposed model from this forecast is that the proposed model may tend to reduce the amount of rainfall, thereby the predicted precipitation could be less than the actual.

C. Computation Time Cost

The total training time for VDSD based on 22 years ACCESS-S forecast data is around 14.7 hours.

Predicting a super resolution climate image needs about 2 seconds per images, and totally 0.13 hour is needed for one-year downscaling, which is faster than traditional methods like ECPP and QM according to table I, but slower than VDSD because more weather variables are used in the proposed model.

TABLE I: Computation time of four downscaling methods (hours)

Method	Training time	Operation time
BI	0	0.02
QM	0	11.21
VDSR	13.12	0.06
VDSD	14.76	1.12

QM doesn't need training, and our implementation took about 0.3 seconds for each station. It needs about 11.21 hours to downscale all the 316*376 grid points for the whole Australia. Another downscaling technique ECPP needed about one hour for training for each station or grid point and took 0.46 seconds for operational forecast [21]. It takes about 15.1 hours to downscale rainfall for the whole Australia.

V. Conclusion and Discussions

Years 2010 and 2011 were 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 La Niña event peak [1].

Year 2012 is in the 2010–12 La La Niña event, the 2011–12 peak was weaker, but still of moderate strength, in both atmospheric and oceanic indicators (http://www.bom.gov.au/climate/enso/lnlist/).

From the above results, our model is better than the interpolation model in statistical results and has a similar performance to the calibration model. From a regional perspective, our model performs better on the west coast of Australia than on the east coast. In terms of time efficiency, our model is a huge improvement, which confirms that usability of it.

However, the theoretical basis of our model is the super-resolution technique, so the results are heavily dependent on the input data. In addition, the uncertainty of the data set itself makes our model more difficult to predict accurately.

There are several directions to move the proposed technique for daily operation in future. Station-based precipitation observations have not assimilated in BARRA-R and its grid precipitation may be not very consistent with on-the-ground observations [25]. To remove such inconsistency, station-specific downscaling techniques like QM or ECPP may further improve long lead time forecasts.

Besides precipitation and geopotential height, we attempted to include other climate variables. An example is including temperature from ACCESS-S1 as a third input image. Compared to VDSD with geopotential height and precipitation, adding the third channel temperature will slightly reduce the skill, especially when the leading time is small. The possible reason is that there is no clear relationship between temperature and precipitation within four different seasons in a year. We may investigate a separate downscaling model for different seasons in future.

For a fair comparison and reducing training time, we only used the forecasts with lead time less than seven days as

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

training data. That may lead to put more emphasis on low resolution precipitation such that no much on correcting inherent bias of GCM's output. 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] K. A. Parton, J. Crean, and P. Hayman, "The value of seasonal climate forecasts for australian agriculture," *Agricultural Systems*, vol. 174, pp. 1–10, 2019.

[3] The Centre for International Economics, "Analysis of the benefits of improved seasonal climate forecasting for agriculture," 2014. Accessed in Nov 2020.

[4] 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.

[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," *KDD'17: Proceedings of the 23rd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 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] 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.

[8] 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.

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

[10] A. Dai, R. M. Rasmussen, K. Ikeda, and C. Liu, "A new approach to construct representative future forcing data for dynamic downscaling," *Climate Dynamics*, vol. 55, no. 1, pp. 315–323, 2020.

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

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

[13] 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.

[14] H. J. Fowler, S. Blenkinsop, and C. Tebaldi, "Linking climate change modelling to impacts studies: recent advances in downscaling techniques for hydrological modelling," *International Journal of Climatology: A Journal of the Royal Meteorological Society*, vol. 27, no. 12, pp. 1547–1578, 2007.

[15] 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.

[16] 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.

[17] L. Chuyao, 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*, 2020. The fine-grained spatial appearances and the position misalignment issue.

[18] 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.

[19] 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.

[20] 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: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, vol. 132, no. 621C, pp. 2925–2942, 2006.

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

[22] 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.

[23] 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.

[24] 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.

[25] 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.

[26] T. Davies, M. J. Cullen, A. J. Malcolm, M. Mawson, A. Staniforth, A. White, and N. Wood, "A new dynamical core for the met office's global and regional modelling of the atmosphere," *Quarterly Journal of the Royal Meteorological Society: A journal of the atmospheric sciences, applied meteorology and physical oceanography*, vol. 131, no. 608, pp. 1759–1782, 2005.

[27] R. Keys, "Cubic convolution interpolation for digital image processing," *IEEE transactions on acoustics, speech, and signal processing*, vol. 29, no. 6, pp. 1153–1160, 1981.

[28] D. Maraun, "Bias correction, quantile mapping, and downscaling: Revisiting the inflation issue," *Journal of Climate*, vol. 26, no. 6, pp. 2137–2143, 2013.

[29] D. Hudson, A. G. Marshall, Y. H. Yin, O. Alves, and H. H. Hendon, "Improving intraseasonal prediction with a new ensemble generation strategy," *Monthly Weather Review*, vol. 141, no. 12, pp. 4429–4449, 2013.

[30] Z. Wang, J. Chen, and S. C. Hoi, "Deep learning for image super-resolution: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[31] J. Kim, J. Kwon Lee, and K. Mu Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1646–1654, 2016.

[32] R. A. Minzner, C. Reber, L. Jacchia, F. Huang, A. Cole, A. Kantor, T. Keneshea, S. Zimmerman, and J. Forbes, "Defining constants, equations, and abbreviated tables of the 1975 us standard atmosphere," 1976.

[33] 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.

[34] M. Xu, Q. Liu, D. Sha, M. Yu, D. Q. Duffy, W. M. Putman, M. Carroll, T. Lee, and C. Yang, "PreciPatch: A dictionary-based precipitation downscaling method," *Remote Sensing*, vol. 12, no. 6, p. 1030, 2020.

[35] Bureau National Operations Centre, "Operational implementation of access-s1 forecast post processing (sep 2019)," Tech. Rep. 124, Bureau of Meteorology, Sep 2019.