EGGS-WG: An open source global gridded stochastic weather generator derived from ERA5-Land

Alex Schuddeboom[1], Christian Zammit[2], David Plew[2], Piet Verburg[3], and Aidin Jabbari[2]

[1]School of Physical and Chemical Sciences, University of Canterbury, Christchurch, New Zealand
[2]National Institute of Water and Atmospheric Research, Wellington, New Zealand
[3]School of Geography, Environment and Earth Sciences, Victoria University of Wellington

Correspondence: Alex Schuddeboom (Alex.schuddeboom@canterbury.ac.nz)

Abstract. A new open source and freely available stochastic weather generator, the **E**RA5-land **G**lobal **G**ridded **S**tochastic **W**eather **G**enerator (EGGS-WG), is publicly available. This model offers several advancements over existing alternative stochastic weather generators, including the ability to simulate any terrestrial region of the planet, moving from simulating a single site to an entire gridded domain and increasing the temporal resolution of temperature simulation from daily to hourly. EGGS-WG is designed to be as low resource and portable as possible, ensuring ease of use for any potential end user. This paper describes in detail all of the principles that are used in the model design and code implementation. Preliminary tests are performed over five climatically distinct regions to confirm the ability of the model to reproduce the behavior of the ERA5-Land dataset. Precipitation occurrence rates and daily rainfall amounts are shown to be reproduced accurately by the model with a range of attributes investigated, including seasonality, spatial correlations and rainfall spells. However, some clear issues remain in the simulation of the most extreme precipitation values. Similar analysis is also performed for the air temperature and dew point temperature, showing stronger agreement than the precipitation data. EGGS-WG is designed with a focus on potential future developments, including support for CMIP6 driven future warming and expansion into simulating other variables provided by the ERA5-Land dataset.

1 Introduction

Stochastic Weather Generators (SWGs) are a commonly used type of model, particularly for the simulation of realistic long term sequences of weather and to generate large ensemble probabilistic forecasts (**????**). The fundamental principle behind SWGs is that by representing weather with a series of statistical relations, we can develop relatively simple and low computational cost models which are able to accurately simulate weather. As these models neglect the physics and chemistry of the earth system, they are fundamentally limited in their capabilities. For example, they struggle to extrapolate beyond historically observed data. However, they have been found to have a wide range of applications due to their unique advantages. This includes, but is not limited to, simulating precipitation for driving complex ecological models (**?**), as a source of multiple variables for crop models (**?**), driving hydrological models including stream flow forecasts (**?**) and acting as an intermediary model to downscale results from global climate models (**?**).

The majority, but not all, of SWGs follow the design principles that guided the initial development of the Weather Generator (WGEN) model (**???**). SWGs of this kind, often known as Richardson style models, are used to simulate precipitation, temperature, winds and radiation. This simulation process can be summarised as follows. First, a Markov chain is used to determine if an individual day is going to be wet or dry. Second, if the day is determined to be wet, then the precipitation amount is simulated by using a random number generator in conjunction with an assumed precipitation distribution (although sometimes an empirical distribution is used). Finally, the other required variables are calculated using a first-order auto-regressive model with lag-1 auto-correlation and cross-correlation between all variables (**?**). The values used in this last step are conditioned by the precipitation state to ensure that any relationships between precipitation and other variables are appropriately represented in the model. Prominent models based on this general approach include WGEN (**?**), CLIMGEN (**?**), CLIGEN (**?**), WeaGETS (**?**) and LARSWG (**?**). There are also more modern models which have a similar design basis but include more sophisticated statistical approaches, such as weatherGEN (**?**), RMAWGEN (**?**) and GWGEN (**?**). An interesting chain of models were also developed over California combining weather states, resampling and Gaussian noise (**????**). There are other kinds of weather generators that fill a similar role but use completely different simulation processes, such as physical modeling (**?**), alternative statistical approaches (**??**) or machine learning (**??**), however, these alternative approaches remain either more computationally expensive, complex or otherwise limited in interpretability, consistency or have to be built specifically for the region simulated.

One of the major limitations of many of these SWGs is that they are restricted to the simulation of a single location at a given time. This is because these initial designs were based around extending limited time series corresponding to weather stations. There has been a long research history working on the development of Multi-site generators (**??????**). Changing from a single-site to a multi-site generator is a complex problem, as simulations additionally need to account for spatial correlations. For instance, if an anomalously hot temperature is simulated at one location, it would be expected that a second location within a few kilometers would also be hot. The most popular approach for handling this problem is the usage of spatially correlated random numbers. In particular, the technique described in **?** and refined in **?** has proven extremely effective at generating the matrices needed to enforce appropriate correlations between any random numbers. There are alternative approaches for the generation of coherent random numbers such as those used in **??**, however, the performance advantages in computational cost or accuracy are relatively minor and in the case of our model would require a much more complex implementation that may not be as consistent over the entire planet.

The motivation for this project was to further these developments with the release of an open source and easy to use gridded SWG capable of simulating any region of the planet, the **E**RA5-land **G**lobal **G**ridded **S**tochastic **W**eather **G**enerator (EGGS-WG). Primarily, this is because there is no existing open-source model of this kind that can be used globally. The aim is to produce the two most important fields for end-users, precipitation and temperature. A particular focus has been placed on ensuring that the resource requirements are minimized, with the aim of reasonably sized simulations taking less than a day on personal computers. The approaches used in **?** were shown to be highly effective at generating a gridded SWG over Canada using high resolution data from the Canadian Precipitation Analysis system. The primary aim of our model was to refine this approach and then use the ERA-5 Land dataset (**??**) as the basis for our model parameters. This results in a model that can create versatile gridded output by automatically calculating the parameters needed to run the simulation. This lets our model

change from the regionally constrained approach in **?** to a global approach substantially increases the usability of the model. The addition of dew point temperature and relative humidity simulations will also be useful for many end users.

It is important to note that while many of the developments of this model were inspired by **?**, the other freely available global models are generally limited. Compared to currently available global models, moving to a gridded approach from a site based approach is a notable improvement. This could be particularly useful for hydrological analysis in which an entire catchment will need to be considered. Additionally, ERA5-Land provides global land coverage, meaning that it should always be possible to estimate our precipitation distribution parameters without complex estimation strategies such as those used to increase the coverage of CLIGEN (**??**) and will not be constrained to a single region (**???**). There are some limitations for ERA-5 data, particularly in the representation of extremes, however there are no viable alternatives to reanalysis data in terms of coverage and resolution. As we show later our model also fails to accurately handle extremes, meaning that we have ~~compounding biases in~~ two indepentent sources of errors in our simulation of extreme precipitation. Additionally, the hourly temporal resolution and high spatial resolution provided by ERA5-Land dataset can produce a level of detail missing from other publicly available SWG with lower resolutions. Despite the clear benefits to doing so, there are very few SWGs that simulate temperature hourly, as is done in **?**, with many of the older models simulating daily minimum and maximum temperatures instead. There are other approaches like that used in the **???** that might be more powerful but their dependence on local weather states makes it infeasible to apply these approaches globally.

The remainder of this paper describes the underlying mathematical details of EGGS-WG, discusses some of the particulars of the implementation, and performs some simple validation tests on output generated. The full code base for EGGS-WG is available on a GPL-3.0 license. This is accessible through a public GitHub, with a version of record tied to this paper released on Zenodo, (https://doi.org/10.5281/zenodo.19228384, **?**). Also included on this repository are instructions for operating the SWG and downloading the data needed to run the model. The only files required to run the model, in addition to the code base, are ERA5-Land data. Initially the EGGS-WG is released with the ability to simulate precipitation, temperature and relative humidity, however other variables, including winds and radiation, are planned to be introduced with further development. As the later results show, performance from EGGS-WG is generally strong across most metrics although clear issues were identified for particular problems that warrant some caution for users. Most prominently amongst these are problems simulating the most extreme precipitation values, slightly limited dew point temperature variability and an underestimation of air and dew point temperature correlation. Paricular issues are also identified in our New Guinea test region.

2 Model Design

As stated in section **??**, the design of EGGS-WG is heavily based on the work of **?**. This general model flow is a slight modification of the Richardson style stochastic weather generators. It has the following three-step structure, with each step explored in detail in a following subsection:

1. A first order two state Markov chain is used to determine the precipitation state for each of the grid cells in the simulation domain. To ensure grid cohesion, the random numbers driving these Markov chains are all required to be spatially correlated.

2. For every cell that is determined to be precipitating, a rainfall amount is calculated using a precipitation equation determined specifically for each location. As with the occurrence, the random numbers used are required to be spatially correlated.

3. Hourly air temperature and dew point temperature are simulated using a modified residual approach. This is different from traditional Richardson generators, as the residuals are generated through the usage of conditioned Fourier series.

2.1 Precipitation Occurrence

The simulation of precipitation occurrence using a Markov chain is one of the longest standing and most consistent component of SWGs (**??**). There has been significant experimentation with different kinds of Markov chains, including using second and higher order chains (**??**), using more than two precipitation states (**??**) and using more sophisticated Markov models to account for spatial correlations (**?**). A simple two state first order Markov chain is used in this work, as they generally provide adequate performance (**?**) and more importantly should be more robust for simulating low precipitation regions. However, in certain climates a single order chain has been shown to fail to produce sufficient dry spells (**???**), as such an exploration of possible improved performance through a more complex Markov chain is intended for the future.

The exact simulation process for the vector of precipitation states, *Q*, is as follows. Every grid cell, *i*, is considered either to be wet or dry. A dry day is defined as a day in which less than 1 mm of precipitation is recorded, with all other days considered wet. The vector is filled so that $Q_{i,t} = 1$ indicates that a grid cell, i, at time t is wet, and a value of 0 is used if i is dry at time t. The simulation is driven using two different probabilities, the probability that a dry day transitions to a wet day, $P_{[Q_{i,t-1}=0,Q_{i,t}=1]}$ and the probability of a wet day remaining wet for the next day, $P_{[Q_{i,t-1}=1,Q_{i,t}=1]}$. For each day a uniformly distributed random number, labelled w_i^t, is generated and if it is greater than the probability of a wet day corresponding to the current state, $P_{c,i}$, the day is considered dry. If w_i^t is less than, $P_{c,i}$ the day is considered wet. In its entirety, this process works by initializing Q to a random precipitation state (more details in section **??**) and then simulating the next state day by day based on the transition probability of the grid cell. To ensure that the resulting field is spatially coherent, the random numbers, w_i^t, used with the transition probabilities must be spatially correlated, as described in further detail in section **??**. To enforce seasonality in the model, $P_{c,i}$ is calculated independently for each month using all the historic data corresponding to that month.

2.2 Precipitation Amount

Due to the statistical complexities of the precipitation distribution, many different models for precipitation amount simulation have been used in past stochastic weather generator research. Most prominent approaches amongst existing models assume precipitation follows an exponential (**?**), mixed exponential (**??**) or gamma (**?**) distribution. Of these approaches the gamma

distribution has the most success at replicating the nature of precipitation, being able to do so from very simple physically driven stochastic models (**?**) and has the strongest case for modelling evolution under climate change (**?**). There are also approaches that use an empirical or semi-empirical distribution (**?**); however, we avoided an empirical approach due to concerns it would limit future development of EGGS-WG, particularly if future climate changes are considered. For EGGS-WG, we choose to use a three parameter mixed exponential distribution following the work of **?**. This choice was due to the relative simplicity of simulating the distribution, the percieved ease of fitting the parameters over a range of climates and the assumption it would be the easiest distrubution to modify for future climates. However, during the review process we were made aware of appropriate techniques for ~~modelling gamma functions~~ using gamma distributions and we aim to update the model representation of precipitation to one based on gamma distributions in a future release. The three parameter mixed exponential distribution can be described as:

$$f(\alpha, \beta_1, \beta_2) = \frac{\alpha}{\beta_1} e^{(\frac{-r}{\beta_1})} + \frac{1-\alpha}{\beta_2} e^{(\frac{-r}{\beta_2})}, \tag{1}$$

where r is the observed precipitation amount, α effectively represents the probability that the exponential associated with β_1 represents the precipitation distribution. Therefore, $1-\alpha$ represents the probability of the β_2 distribution. This also requires the enforcement of the bounds $0 < \alpha \leq 1$ and $\beta_1 \geq \beta_2 > 0$. The parameters, α, β_1 and β_2, are calculated independently for every grid cell in the simulation region. Once we have fit these parameters, an inverse CDF (cumulative distribution function) procedure can be used to simulate the precipitation amounts. Following **?**, this can be done with the equation:

$$r_i^t = DLT - \beta_i^t \times ln[1 - \Phi(z_i^t)], \tag{2}$$

where r_i^t is the rainfall that occurs at grid cell i and time t. DLT is the minimum viable precipitation required for a day to be considered wet (1 mm/day for EGGS-WG), and is required because imposing this cutoff on our simulation means that values below 1 mm should not exist in simulation output. z_i^t is the random number (generated from a standard normal distribution) used to drive the simulation. $\Phi(\cdot)$ is the function transforming a Gaussian random number to a uniformly distributed equivalent. β_i^t is a combination of β_1 and β_2:

$$\beta_i^t = \beta_{2,i} + 2(\beta_{1,i} - \beta_{2,i})[1 - \Phi(w_l^t)/\alpha_i^t P_{c,i}] \tag{3}$$

This is following the approach introduced in **?**, which uses the occurrence model data as part of determining β_i^t. This includes both the transition probability, $P_{c,i}$, and driving random number, w_i^t from the occurrence simulation. A combined model approach like this is essential for avoiding unrealistic precipitation fields in the synchronized fields. Both w_i^t and z_i^t must be generated in a spatially correlated way to ensure that the ~~output the~~ model produces spatially coherent output. As with the precipitation occurrence, all model parameters are fit seperately for each gridcell and month to ensure good representation of the seasonal cycle.

2.3 Air Temperature and Dew Point Temperature

By opting to simulate temperature data at an hourly temporal resolution, EGGS-WG follows a different approach than the majority of publicly available stochastic weather generators. Specifically, the approach used here is to generate both a daily and an hourly Fourier series. The hourly series is independent of precipitation but is conditioned by the daily series, enabling the model to account for precipitation state. These Fourier series are used to reduce the underlying data to normally distributed residuals, which in turn are used to determine the parameters necessary for simulation. Fourier series are particularly useful for representing temperature due to the large periodic signals due to seasonal and diurnal variations. As far as the authors are aware, this exact approach has only been used by **?** to date; however, they recognize that their approach is partially based on other existing work (**??**). First, the hourly Fourier series is calculated using the entirety of the input data. This is done for every grid point i using the equation:

$$\tilde{P}_H(i,t) = \bar{T}(i) + \sum_n \{\alpha_H(i,n)cos[\frac{2\pi nt}{8760}] + \beta_H(i,n)sin[\frac{2\pi nt}{8760}]\}, \tag{4}$$

where $\tilde{P}_H(i,t)$ is our function that represents the Fourier series fit for either the mean or standard deviation of the hourly temperature data at location i and time t. The series is calculated for both mean and standard deviation independently, as both are required to calculate residuals. $\bar{T}(i)$ is the mean temperature or mean standard deviation over the entire record (the variable matches whichever $\tilde{P}_H(i,t)$ variable is being calculated) and α_H and β_H are Fourier coefficients corresponding to a particular harmonic, n. For EGGS-WG, we use the values $n \in [1, 365, 730, 1095, 1460, 2190]$ corresponding to yearly, daily, 12-hourly, 8-hourly, 6-hourly and 4-hourly cycles. This is a possible point of improvement for EGGS-WG, as different selections of n may lead to a more realistic temperature cycle, although initial work has only identified minor impacts of changing these values. α_H and β_H are calculated with the following formulations:

$$\alpha_H(n,i) = \frac{2}{8760}\sum_{t=1}^{8760} \bar{M}_H(t,i)cos(\frac{2\pi nt}{8760}) \tag{5}$$

$$\beta_H(n,i) = \frac{2}{8760}\sum_{t=1}^{8760} \bar{M}_H(t,i)sin(\frac{2\pi nt}{8760}), \tag{6}$$

where $\bar{M}_H(t,i)$ is either the mean or standard deviation (~~again to~~ chosen to match $\tilde{P}_H(i,t)$) over the entire observational record at time of year t and location i. With both the hourly mean, $\tilde{\mu}_H(t,i)$, and standard deviation, $\tilde{\sigma}_H(t,i)$, calculated using equation **??**, a full set of residuals can be calculated by subtracting $\tilde{\mu}_H(t,i)$ and dividing by $\tilde{\sigma}_H(t,i)$ from the original temperature data. To avoid confusion~~between each of these different kinds means we clarify~~, $\bar{T}(i)$ is the mean value over the entire time at location i, $\bar{M}_H(t,i)$ is specific to time of year t and location i and $\tilde{\mu}_H(t,i)$ is the mean generated by our Fourier series for time of year t and location i. Once the data is processed into hourly residuals, they can be used to determine two matrices, $\mathbf{X}$ and $\mathbf{Y}$, needed to run an autoregressive model (**?**). In particular, this model takes the form:

$$\hat{R}_H(t) = \mathbf{X}\hat{R}_H(t-1) + \mathbf{Y}E(t)\,, \tag{7}$$

where $R_H(t)$ is the series of the residuals that can be converted back to temperatures using the Fourier series. As is clear from equation **??**, there are two components to the model: one part dependent on the last hour residuals, $\hat{R}_H(t-1)$, and one part dependent on $E(t)$, which is a vector of randomly generated normal variates with a mean of 0 and standard deviation of 1. Calculation of **X** and **Y** is a well established process (**???**), using the unlagged, C_0 and lag-1, C_1, correlation matrices between residuals. This lets us calculate **X** and **Y**, with:

$$\mathbf{X} = C_1 C_0^{-1} \tag{8}$$

$$\mathbf{G} = C_0 - \mathbf{X}C_1^T\,, \tag{9}$$

where **G** is the Gramian of **Y**, so **Y** can be calculated with the Cholesky decomposition of **G**. With **X** and **Y** calculated, we have everything needed to run our hourly temperature model in equation **??**. However, as mentioned above, this process leaves the generated temperature data independent of the precipitation state. To condition these fields, we generate two different daily Fourier series, one using data from only wet days and one with only dry days. Again, we follow the approach detailed in, **?** which uses the following modified versions of equations **??**, **??** and **??**:

$$\tilde{P}_D^s(i,t) = \bar{T}^s(i) + \sum_n \{\alpha_D^s(i,n)cos[\frac{2\pi nt}{365}] + \beta_D^s(i,n)sin[\frac{2\pi nt}{365}]\} \tag{10}$$

$$\alpha_D^s(n,i) = \frac{2}{365}\sum_{t=1}^{365} \bar{M}_D^s(t,i)cos(\frac{2\pi nt}{365}) \tag{11}$$

$$\beta_D^s(n,i) = \frac{2}{365}\sum_{t=1}^{365} \bar{M}_D^s(t,i)sin(\frac{2\pi nt}{365})\,, \tag{12}$$

where S is the precipitation state, and is used to indicate the variable is only using data specific to that precipitation state. An alternative set of n values are used for this calculation as well, $n \in [1,2,3,4,6]$, to account for the shift from hourly to daily. The difference between the daily averages from the unconditioned series from equation **??** and the conditioned series from equation **??** is calculated and used to adjust the hourly series, resulting in two conditioned hourly series that can be used to transform simulated residuals to conditioned temperature values. This is simply done by adding the calculated difference to the hourly series, bringing it in line with the daily series but retaining the hourly structure. This entire simulation process can be summarized as follows:

1. An hourly Fourier series for both mean values and standard deviation is fit to the underlying air temperature and dew point temperature data. This process is done independently for every grid cell in the domain.

2. Distinct dry and wet daily Fourier series are calculated using only data from dry or wet days for the mean and standard deviation of both variables.

3. Each hourly Fourier series is split into a specific wet and dry series by calculating the difference between the daily mean of the hourly series and the corresponding daily series and then adding that difference to the hourly series.

4. The conditioned series mean and standard deviation values are used to transform the existing data into residuals, which in turn are used to calculate the **X** and **Y** matrices required to drive the model.

3 Implementation

While section **??** describes the design principles behind EGGS-WG, there are several implementation level issues that need to be further explained to understand how the model functions. Several of these decisions were made primarily from practical concerns with a view towards limiting the resources required to run EGGS-WG, and building EGGS-WG in a way that does not restrict future development. The most substantial of these issues are discussed in the following subsections.

3.1 ERA-5 Land Data

All the parameters required to run the model are derived from ERA 5 Land output (**?**). This is a $0.1^\circ \times 0.1^\circ$ spatial resolution dataset with hourly values and a global coverage over land surface. Specifically, the variables used for this model include the 2 m air temperature, 2 m dew point temperature, precipitation and geopotential (used to determine altitude which is output but not used for simulation). The ERA5 download process allows for the data to be subset before being downloaded to disk. As such, disk space and memory costs can be reduced by using a subset covering the simulation domain. Code to download this data is included with the model. Expansion to other ERA5-Land variables including winds and radiation is planned but not included in this version of EGGS-WG. Other potential additions include surface heat fluxes, evapotranspiration and soil moisture content, but questions remain about feasibility. For model validation purposes, the ERA5-Land data is referred to as observational data in this paper, although as ERA5 is a reanalysis product this is not strictly true. In general, ERA5 has been shown to represent precipitation well in comparison to gauge data with a slight wet bias, however the most extreme precipitation values appear to be underestimated (**?**) which will impact our simulation of the these extreme values.

3.2 Stationarity

Most stochastic weather generators are inherently stationary (**?**), meaning that the climate simulated is fixed to one point in time and effects such as past climatic changes due to greenhouse gases are unaccounted for. There is, however, an established approach adding climate change support to SWGs and using them for down scaling larger resolution climate models (**?????**). While we are interested in introducing this capability to EGGS-WG at a later date, currently EGGS-WG is designed to be

stationary. To account for the historical changes that have occurred over the observational period, we adjust all the ERA5-Land data so that it is equivalent to data from the year 2022. This means that our SWG is only effectively simulating conditions equivalent to 2022. To adjust our data, we fit a Fourier series to the underlying data. The linear terms of this Fourier series are then used to generate a trend line for the entire observational record. This trend line can then be inverted and added to the observational data, adjusting the data so that it matches 2022 climate. For example if June showed 1 degree warming between 1950 and 2022, then 1 degree would be added to the 1950 values. Only the linear term is used so that seasonal variations remain in the data, this may seem an unusual choice but as our model effectively works with month specific data we don't need a seasonal correction. This is why a simple Fourier series approach is suitable for this problem. Adjustments made using this approach are limited and will not account for some potentially significant changes in variability or precipitation occurrence.

Additionally, the model has been released with a non-stationary mode that is not examined in this paper. Effectively, this non-stationary mode adds the trends observed in the Fourier series back into the model output after the simulation process. This results in the model recreating the observed trends in ERA5-land without any fundamental changes in the simulation calculations. Periods from before the ERA5-land record are all simulated at 1950 conditions, as there is no consistent approach for estimating earlier changes with the ERA5-Land data. Periods from beyond ERA5-land coverage are simulated with the most recently available data, assuming no change into the future. As stated in the above paragraph, climate change forcings will be considered in the future, but this will go beyond the data provided by ERA5-land and utilize the CMIP model output. An example figure showing the effects of these adjustments on temperature is included in the supporting information as figure S1. Given the same adjustment is made to observations and simulations, it should have no impact on our model validation, provided we are comparing data of the same kind.

3.3 Precipitation Parameter Calculation

There are several distinct approaches that could be used to solve for α, β_1 and β_2 in equation **??**. In EGGS-WG these variables are estimated by using the SciPy optimization package (**?**) to minimize the difference between the theoretical distribution in equation **??** and the ERA5-Land data. Specifically, this is done by using the SciPy *optimize.minimize* to minimize the negative log-likelihood of this function. This calculation is done independently for every grid point and every month. These values need to be calculable for every grid point in the ERA5-Land dataset, as any grid point could be a potential part of a simulation. This strongly limits the ability to make any simplifying assumptions about the fit of the precipitation function. Experimentation with different algorithms has shown the most consistent results from the Powell (**?**) and Nelder-Mead (**?**) algorithms. The Nelder-Mead algorithm is tried first, and when that fails to reach an appropriate solution, the Powell algorithm used. Failure is determined by the *optimize.minimize* function returns and if neither minimization works the code is stopped and user notified. Restricting α to a value between ~~0.5~~ 0.1 and 0.9 has also shown to improve consistency and computational efficiency in our internal testing. This is because ~~restricting α from 0.5 to 1 just locks in the first exponential as the leading term and moving the upper limit from 1 to 0.9 ensures the solution does not reduce to~~ it prevented the solver from collapsing to a single exponential ~~(although the betas can still be equal as a valid solution)~~case. This was paired with a restriction on β_1 and β_2 to between 0.1

and 1000 which helped remove some unusual edgecases and improve performance. Although the testing has been limited to the regions discussed in this paper and the universal impacts of this restriction are unclear.

3.4 Spatial Coherence of Random Numbers

The random numbers used for the generation of precipitation occurrence and amounts simulations are calculated with the exact same process. In particular, the approach used by **?** (in turn based on **??**) is the one that is implemented in EGGS-WG. One of the difficulties of this process is that it requires swapping between the underlying uniformly distributed values to normally distributed values to make the simulation process efficient (**??**). This component of our model does differ significantly from **?**, opting instead for the simpler approach of **?** because we believe it will be more consistent over a wide range of environmental conditions.

First, we focus on the precipitation occurrence simulation. This process starts by the calculation of the Pearson correlation coefficients between all station pairs of observed precipitation sequences, $\rho_{i,j} = corr(Q_{i,:}, Q_{j,:})$, where $Q_{i,:}$ is the precipitation occurrence vector (as defined in section **??**) at location i over the entire observational record. This lets us define the matrix C which is populated by $\rho_{i,j}$ for every pair of i and j. Provided that C is positive definite, applying a Cholesky factorization will result in the upper triangular matrix R such that $C = R'R$ (**?**). Multiplying the lower triangular matrix R' by a matrix of random normally distributed numbers results in a matrix with the same covariance as C, therefore recreating the expected spatial correlations.

There are, however, some issues that make the generation of R' more complex (discussed in detail in **?**). Most importantly, the generated correlations will be smaller than those observed. **?** explains that this difference is due to needing a higher correlation between random numbers for them to reproduce the expected correlations with this simulation technique. To account for this, larger correlation values than observed will need to be used to generate R'. Calculating the optimal correlation values to use is done iteratively, with the equation:

$$C_{R_{i+1}} = C_{R_i} + \eta(C_{x,obs} - C_{x,syn}), \tag{13}$$

where $C_{x,obs}$ is the observed correlations defined as C above, $C_{x,syn}$ is the correlations resulting from when a model is run with the correlations C_{R_i}. $C_{R_{i+1}}$ is calculated, and the process is repeated until $C_{x,obs} - C_{x,syn}$ reaches a minimum. The correlations for each month of data are handled independently. In EGGS-WG, several different convergence factors η (in particular [0.4,0.2,0.3,0.05] for occurrence and [0.4,0.25,0.1] for amount) are used based on internal testing and $C_{x,obs}$ is used as C_{R_0}. The model runs are set to different lengths given the domain size, ranging from 3000 to 12000 days for each month, with $C_{x,syn}$ taken as the average of between 5 and 15 simulations of this length. This approach can result in $C_{R_{i+1}}$ matrices which are not positive definite. To ensure we have a positive definite matrix, we adjust this matrix by finding its eigenvalues, replacing any negative eigenvalue with a small positive value (1×10^{-12}) and then recreating the matrix by using the new eigenvalues and old eigenvectors (**?**). The precipitation accumulation simulation can be handled the same way with one key difference, the correlations between points, ρ, can only be calculated when both points are considered to be wet.

3.5 Temperature Cross Correlation

One of the differences between EGGS-WG and the SWG in **?** is that EGGS-WG simulates dew point temperature in addition to air temperature. These two variables are obviously highly correlated, so a mechanism for enforcing this correlation in the model is required. This can be done with the well established approach defined in **??**. Effectively, this is a modification of the matrices **X** and **Y** in equation **??** to include both air temperature and dew point temperature, and changing $\hat{R}_H$ from a vector of air temperature residuals to a vector that includes residuals for both air and dew point temperatures. This requires calculating both the lag 0 and lag 1 cross correlations between air temperature and dew point temperature, with the redefinition of C_0 and C_1 from section **??** as:

$$C_0 = \begin{bmatrix} C_0(T,T) & C_0(T,D) \\ C_0(D,T) & C_0(D,D) \end{bmatrix} \qquad C_1 = \begin{bmatrix} C_1(T,T) & C_1(T,D) \\ C_1(D,T) & C_1(D,D) \end{bmatrix} \tag{14}$$

where $C_0(T,D)$ indicates the lag 0 correlation matrix between air temperature, T, and dew point temperature, D. This approach can be scaled to as many variables as required, with past work commonly including radiation and winds. It is our intention to use this same approach when investigating the addition of new variables in later model releases.

3.6 Vector Initialisations

There are several different approaches for initializing the vectors for simulation. EGGS-WG uses a simplistic approach where a slightly simplified version of the model is run for a spin up period of 1000 days. The first day of this spin up period is considered entirely dry, with the temperatures starting with a residual value of 0. Internal testing showed that this spin up period was sufficient to start the simulation in a random state. For example, in the supporting information figure S2 shows the range of residuals present after 10000 different 1000 day spin up simulations. These residuals clearly follow the theoretical Gaussian distribution very closely showing the expected range of possible starting conditions.

3.7 Post simulation adjustments

A relatively unique innovation of this model is the ability to provide post simulation adjustments to the simulation output. There are many reasons why a user would not want these adjustments on simulation output (mostly due to impacts on future climate simulation), so this is designed to be an easily togglable model feature. Specifically, two series of adjustments are available, a distribution wide precipitation adjustment and a diurnal variation adjustment to temperatures. For the precipitation adjustment, a cubic spline is fit to the differences between the observed and simulated precipitation distributions at specific percentiles. The generated spline is used to calculate an adjustment factor for every simulated precipitation value, functioning as a light quantile mapping correction to the data. This is done using all the precipitation data, so that there is no additional adjustment to account for seasonality or spatial differences. This is also done at a coarse scale, around 10 distinct bins, to avoid overfitting to the underlying data. The impact of this fit can be seen in the supporting information in figure S3, where the general Q-Q plot is mostly unchanged and the data is most significantly altered between 2 and 10 mm. The temperature

adjustment works by examining the difference between the month specific daily mean and the diurnal cycle relative to the mean. These differences are added to the simulated temperature data, adjusting it so that the diurnal cycle in EGGS-WG better matches ~~simulations~~ERA5-Land. The main reason these are included is that the global nature of the model may lead to cases where some model parameters are poorly fit and adjustment is necessary. All figures and analysis in this paper are performed with these adjustments unless otherwise identified.

3.8 Relative Humidity

EGGS-WG does not directly produce the relative humidity through the simulation process. Instead, relative humidity is calculated by taking the ratio of the saturation vapor pressure, p_s, to observed vapor pressure, p. In particular, we use the Tetens equation for estimating vapor pressure with the air temperature T and dew point temperature D. Note that for this equation to be accurate, temperatures must be given in Celsius.

$$RH = \frac{p}{p_s} \qquad p = 0.61078e^{\frac{17.27D}{D+237.3}} \qquad p_s = 0.61078e^{\frac{17.27T}{T+237.3}} \tag{15}$$

4 Results and Discussion

4.1 Simulation Testing Regions

To evaluate the performance of the EGGS-WG, we examine simulations covering three distinct domains over New Zealand, one over Australia and one over New Guinea, as shown in figure **??**. While almost all of New Zealand is the same Köppen type, Cfb, there is substantial variation in the frequency and magnitude of precipitation, particularly when it comes to extreme events (**??**). The three New Zealand based regions are chosen to explore a range of precipitation conditions. The West Coast region of the South Island is a rainforest known for its relatively frequent and high intensity precipitation due to the orographic effects of the nearby Southern alps. On the other side of the Southern alps is the Canterbury region, dominated by flat and dry plains. Finally, we include a region on the North Island around New Zealand's largest freshwater lake, Lake Taupō. This is in-between the other two regions in precipitation frequency and annual rainfall. In addition, we include an Australian region which is amongst the driest environments on the planet and a New Guinea region amongst the wettest. The New Guinea region is also much larger than the other regions, allowing us to test model performance over a large domain. For the observational data, we use trend adjusted ERA5-Land data from the years 1950 through 2022, whereas for the model data simulations of 110 year duration are used.

4.2 Precipitation Testing

The first check for model accuracy is to ensure that the overall precipitation occurrence rates and accumulation values match those generated from observations. Figure **??** shows the observational and simulation values for monthly average occurrence rate and precipitation amount for each of the five regions. Clear differences are seen between the regions with occurrence rates

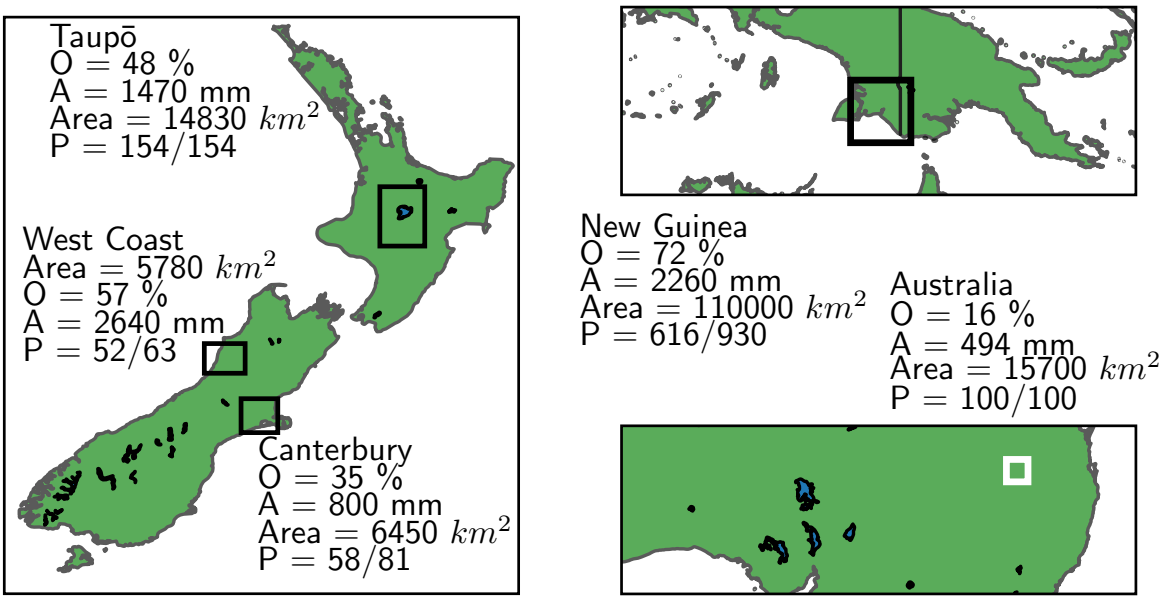


Figure 1. A map showing the five simulation regions in our domain: the West Coast, Canterbury, Taupō, New Guinea and Australia. Included is the ERA5-Land average precipitation occurrence rate over the region, *O*, the annual average precipitation over the region, *A*, the area covered by the region and the total number of grid points in the region, *P*. These values are calculated as averages between 1950 and 2022.

ranging from lows around 10 % over Australia and highs around 100 % in New Guinea and monthly amounts ranging from around 10 mm to 300 mm. Generally, the simulation appears to be able to capture the mean regional properties with some minor differences. All the regions show some level of seasonality, with the monsoon seen in the New Guinea resulting in much larger seasonality than the other regions and the relatively dry Australia and Canterbury regions showing only minor seasonal variation. In general the precipitation occurrence and amount are well captured over all five regions, with the occurrence simulation slightly better matching the observations than the amount simulation.

While figure **??** shows that the domain wide simulation is accurate, it does not give any insight into the spatial structure within these regions. Following an approach established in past research (**???**), the ability of the simulation to capture spatial dynamics was tested using Pearson correlation coefficients. This testing is done by calculating the correlation coefficient between every possible pairing of grid points in the observations and then compared to simulations. All of these calculations are done for each month of the year independently, ensuring that seasonal variation isn't inflating the correlation values. The observed correlation coefficient can then be directly plotted against the simulations to determine if the same spatial patterns of correlation are present.

The resultant correlation coefficient plot is included in figure **??**. Through the simulated correlation coefficients, the simulation is shown to be very successful at replicating spatial precipitation occurrence dynamics over all five of our test regions. Interestingly, the observed correlations are substantially different over the different New Zealand domains with strong cor-

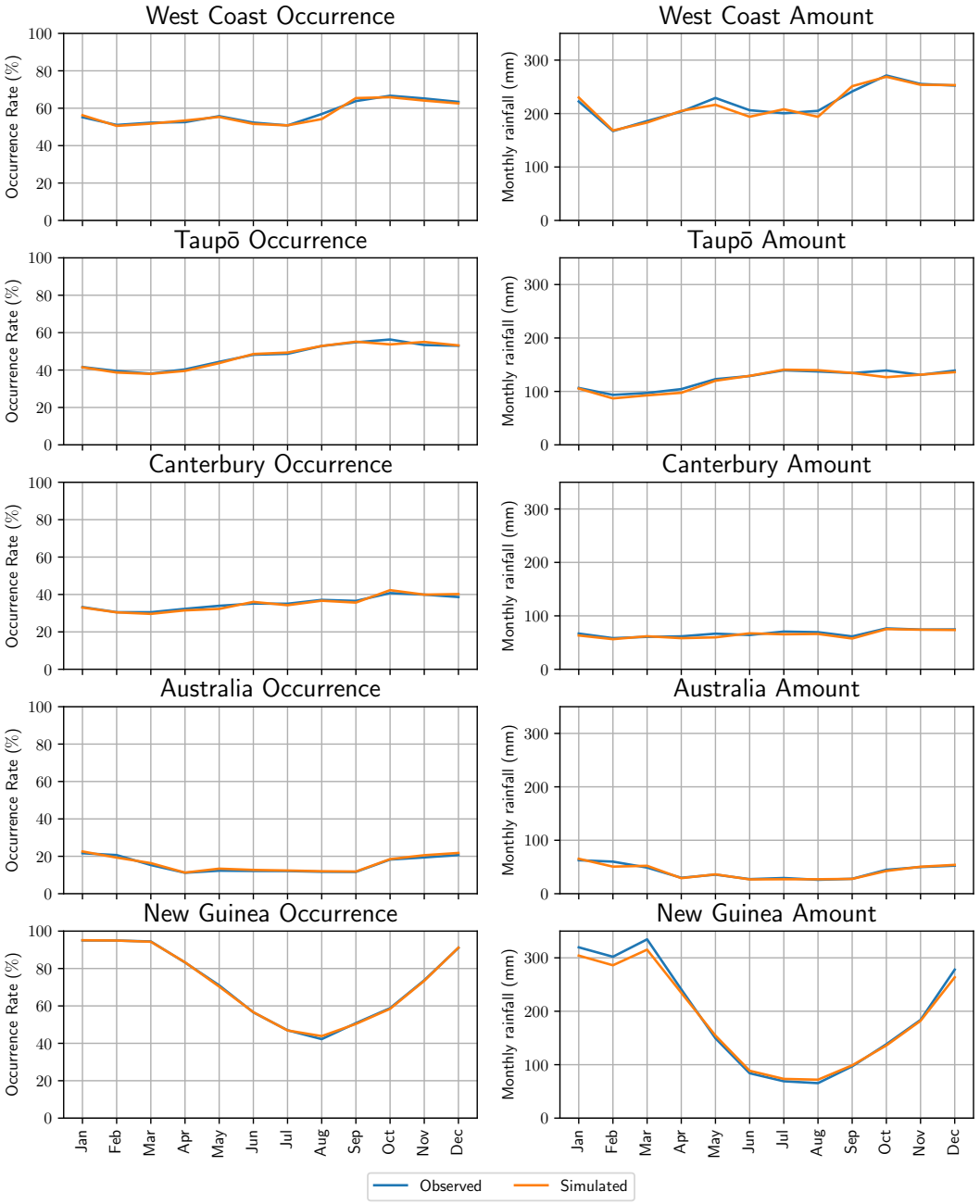


Figure 2. Seasonal average values for the precipitation occurrence and amounts over the five simulation domains. Precipitation occurrence is defined as the percentage of days with more than 1 mm of precipitation. Precipitation amounts are calculated as a monthly total values and then averaged over all the years in the observations and simulations.

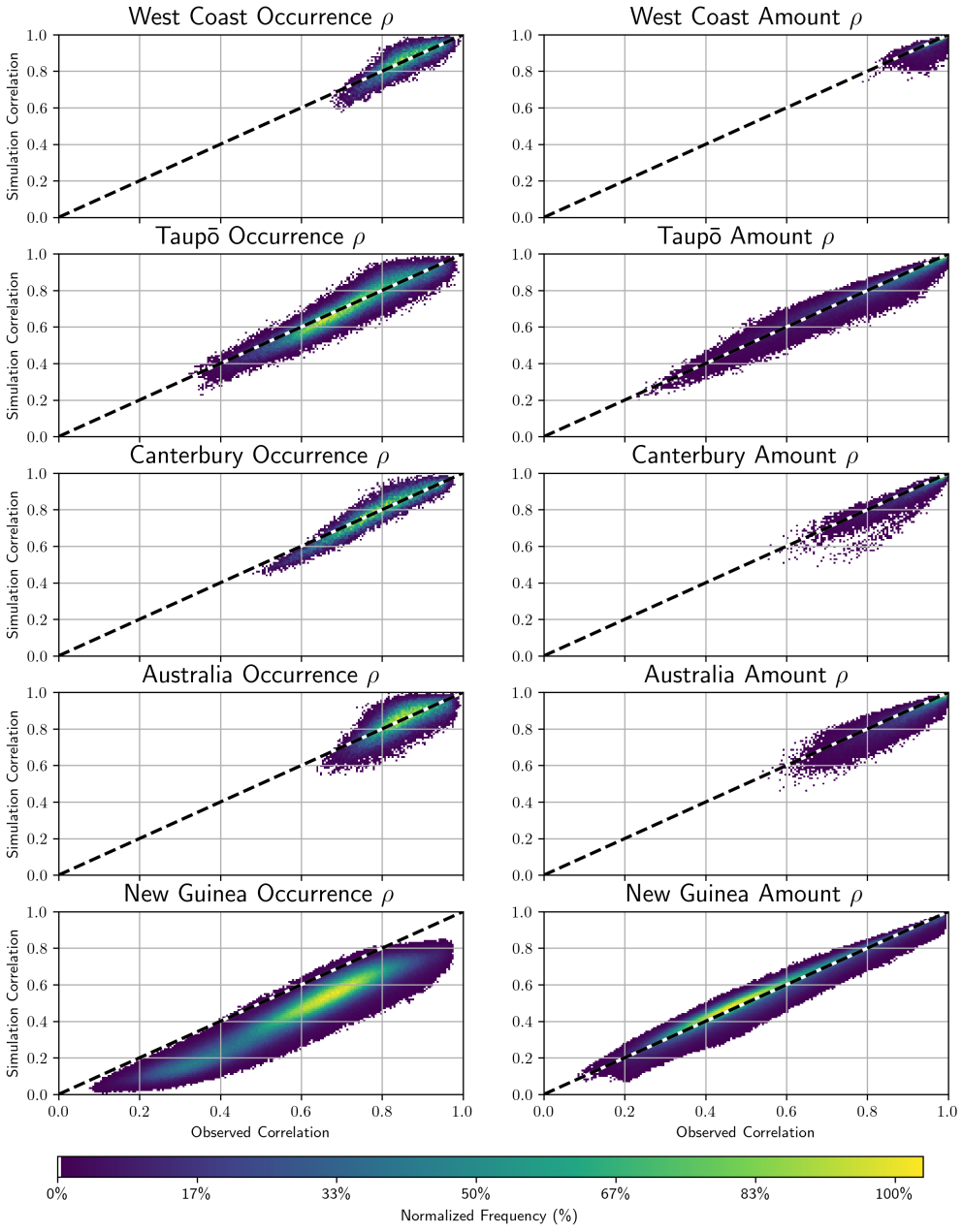


Figure 3. The Pearson correlation coefficient between every set of station pairs for both the simulation and observational data. Values are calculated for both the precipitation occurrence and amount data. The color of the figure indicates the density of station pairs at a given correlation, ranging from purple for a low number of points to green for a high number of points. The included dashed line indicates the 1:1 correlation line that would be the theoretical ideal for a simulation.

relations with tight clustering over the West Coast region, much weaker correlations present over the Taupō region and an intermediary state in the Canterbury region. The simulation is able to capture these dynamics well with a very high level of accuracy in the West Coast and Australia regions, a few outlier grid point pairs that underestimate the observed correlation rate over Canterbury and a wider range but generally accurate values over Taupō. The New Guinea region shows a much different set of correlations than are seen in the other regions, with the main values offset from the 1 to 1 line and points ranging from a correlation of 0 to 1. The spatial dynamics are very similar for the amount values, with the biggest differences seen in the Canterbury region where a relatively small, but clearly visible subset of grid point pairs show much lower simulated correlations than observed. It is worth noting that this a relatively small portion of the simulation pairs over the Canterbury region, and that the overwhelming majority of grid pairs over this region are well captured. The Australia, West Coast and Taupō regions show stronger agreement than the data over Canterbury. The New Guinea amount data shows a better correlation than it does for occurrence, with the majority of the points landing near the 1 to 1 line despite still covering a large range of correlations.

Another common validation approach in past research has been ensuring that precipitation occurrence spells are well represented (**?**), where precipitation spells are defined as consecutive days that are either wet or dry days. These results are included in figure **??** with an alternative version included in the supporting material which uses a logarithmic y-axis to emphasize differences in extremes values. Over all the regions, the agreement between model and observations is strong, although there is minor disagreement, particularly over the New Guinea region. The biggest differences are seen in the short-lived spells of less than 5 days, although typically the biases cancel each other out over the first 5 days. Interestingly, while there are clear distinct behaviors associated with the different simulation regions, such as the relatively short maximum dry day spells on the West Coast or wet day spells in Canterbury, the model is capable of mimicking these distinct regional behaviors. There are some issues with the New Guinea simulation for long streaks, which are visible in the percentiles, sum of absolute errors, and in the supporting information. In fact, the sum of absolute errors shows the biases over the New Guinea region are over twice the size as those over any other region. This warrants further investigation, but given the relative rarity of these events, should not have a significant impact on the simulation. These results lack some of the biases that are expected of using a first order Markov chain, in particular the inability of these models to accurately capture long dry spells (**?**). This ~~may be due~~ appears to be related to the climates of the regions ~~analyzed but does suggest that, at least for these simulation domains, the~~ analysed as some biases in long dry spells are seen specifically in the New Guinea region (further analysis shows this is independent of domain size). Preliminary analysis of additional depth models not included here shows only very minor benefits of changing to a different chain depth~~or design is unlikely to significantly improve EGGS-WG~~, and suggests the issues over New Guinea may require a fundamentally different approach to resolve.

For a closer examination of the precipitation amount distribution, a quantile-quantile plot of the precipitation produced by the simulation and seen in the observations is shown as figure **??**. All the dry day data is removed from this comparison, as this would effectively be testing the occurrence rate of the simulation, not the precipitation amount distribution. Generally, there is very good agreement between the simulated and observed percentiles over all five of the domains with very little difference visible below the 99^{th} percentile. However, the highest percentile values show notable deviation from the rest of the data, with simulation values consistently larger than observed values. This deviation is much more pronounced in the Canterbury

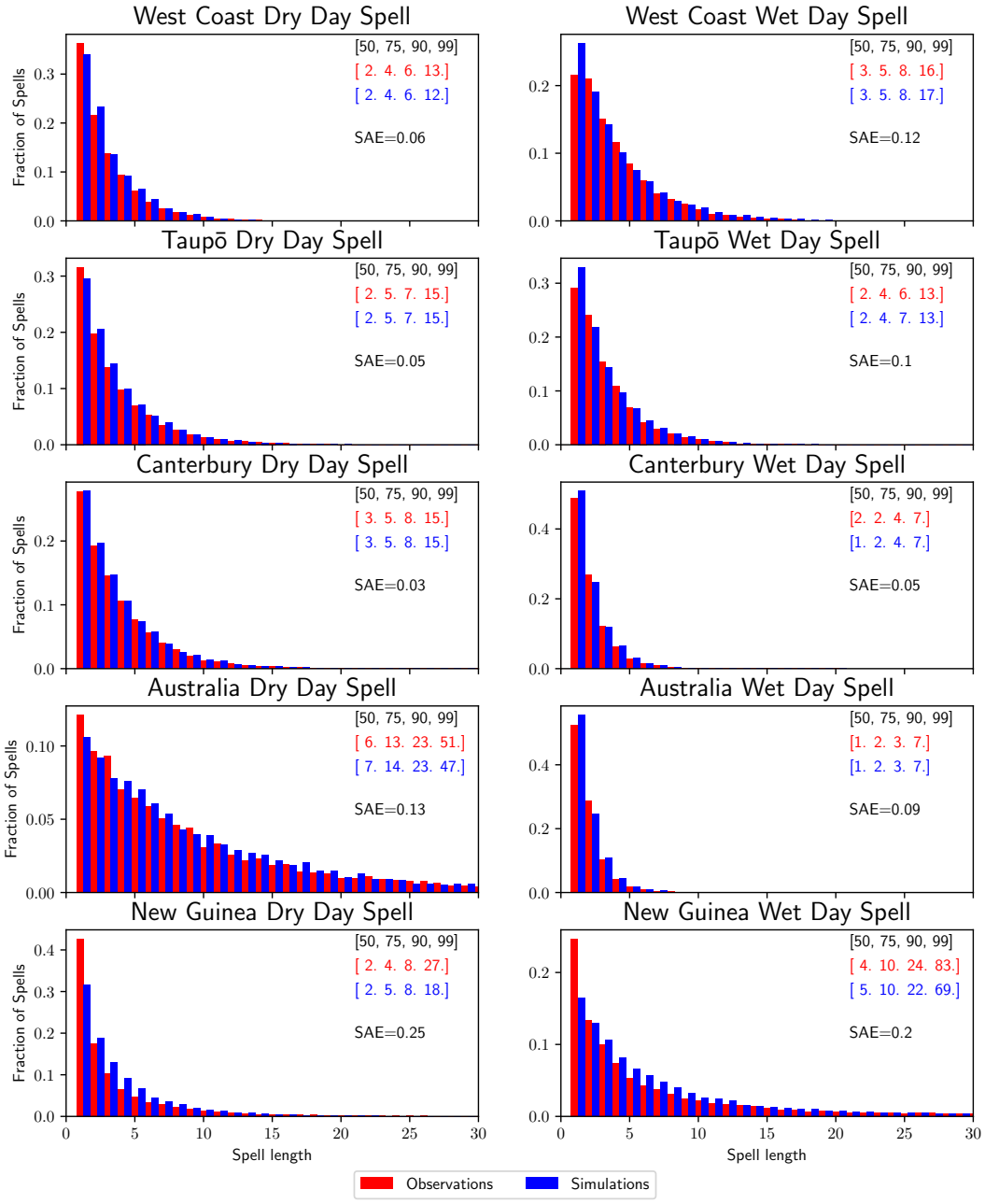


Figure 4. The wet and dry day precipitation spells for the both the simulations and observations over all five of the ~~analyzed~~ analysed domains. A dry spell is defined as number of days in a row below 1mm precipitation where wet is above 1mm. Each bar represents the fraction of total spells that are of a given length. Also included in each subplot is the 50^{th}, 75^{th}, 90^{th}, and 99^{th} percentile (in red for observations and blue for simulations) as well as the sum of the absolute errors (SAE).

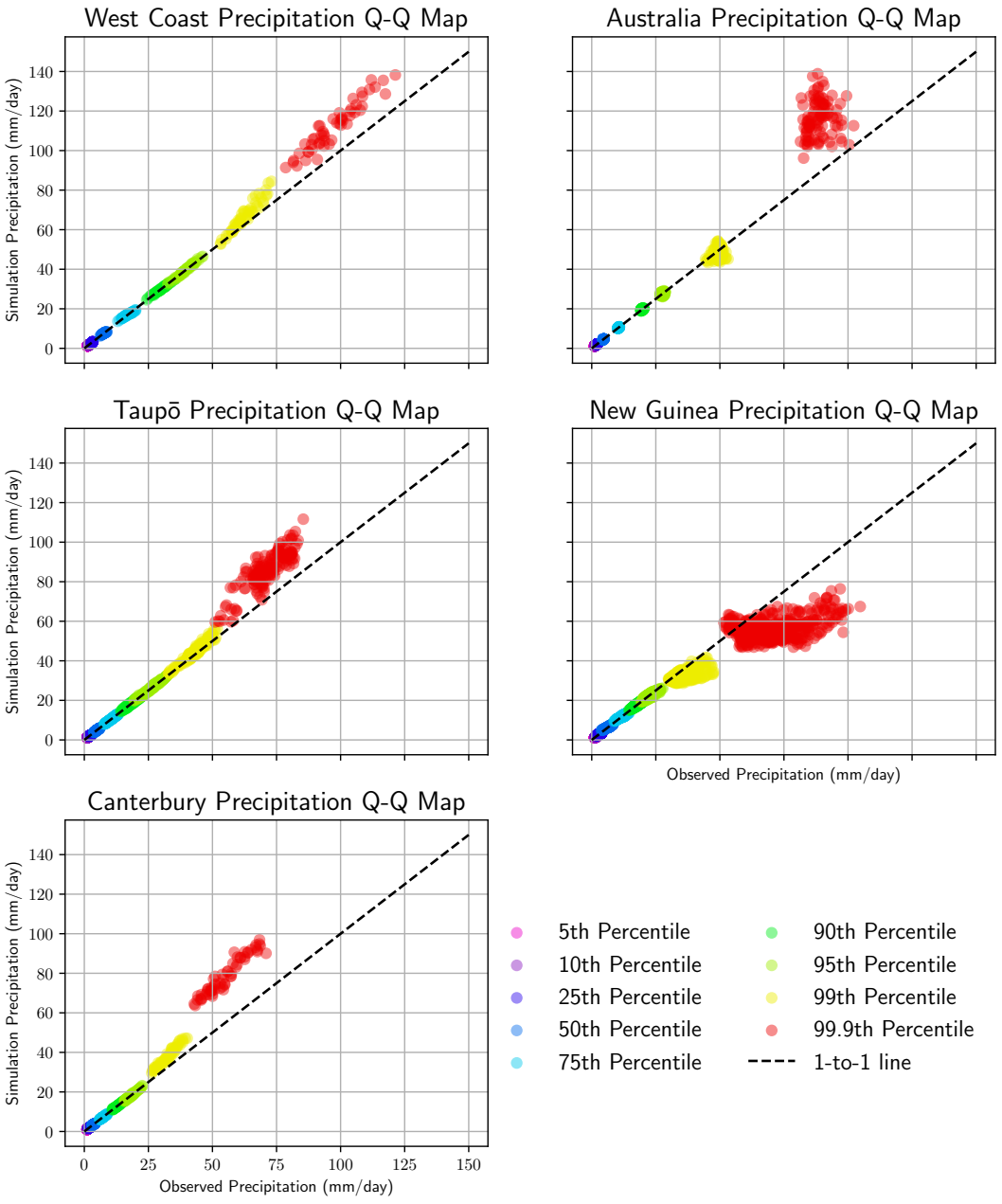


Figure 5. Quantile-Quantile plot of the distributions of observed and simulated precipitation data for every individual grid point in each of the domains. The precipitation data is filtered so that only the wet days are considered. Each dot represents an individual grid point at a corresponding percentile. The percentiles included in this plot are the 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, 95^{th}, 99^{th} and 99.9^{th} percentiles as indicated by the color of the dots. The dashed line indicates the 1:1 correlation line.

region than in the Taupō and West Coast regions and much weaker in the Australia region. The only exception to this is the New Guinea region where the extreme precipitation is underestimated by the simulation. This discrepancy could have major implications, depending on the region, if the model is used to directly study extreme events~~, but for general simulations~~. For general purpose simulations, the impact on the simulations is relatively minor due to the relative infrequency of these events however this could have a substantive impact if studying flood risk or long return period events. A more aggressive precipitation adjustment than the one described in section **??** could be used to improve the simulation of these extreme values. This would likely be a strong quantile mapping correction, but could conceivably also take the form of fitting a different function for when precipitation amounts when values above a certain threshold are reached. However, to retain the simplicity and to avoid overfitting the underlying data, we choose to keep the post simulation adjustments to precipitation relatively small. Additionally, it is important to keep in mind this is a comparison to ERA5 data, which is in turn biased, particularly for extremes, relative to true observational data.

To more comprehensively explore the biases in precipitation distribution, an additional set of figures is included in the supporting information in figures S5-10. These include the general distribution, time aggregated overdispersion plots to align with figures 9 and 10 from **?** and time period aggregated and full region wide distributions. The results shown in the full precipitation distribution show very strong agreement with figure **??**, consistent quality simulation with some minor issues appearing in the representation of the most extreme values. The monthly and yearly standard deviations show results that are comparable to **?** for the majority of the regions. There is however a clear distinct behavior over the New Guinea region where simulated variance is lower than is observed, particularly in the annual cumulative values. The monthly and yearly distributions of the mean values show good agreement but some evidence of underestimated variability is seen in the West Coast and New Guinea regions. Moving to the regionwide mean distribution, there is the emergence of significant New Guinea bias which suggests that the region is not varying as coherently in the model as in the observations, however all of the other regions show good agreement between model and observation. Additionally, the distribution of maximum observed values shows reasonable agreement across the distribution. The regionwide maximum values do however show evidence of an underestimate of the most extreme components of the distribution over the New Guinea region. ~~It is unclear how these regional biases are impacted by the larger size of the New Guinea region, which warrants further investigation but is considered out of scope for this work.~~

4.3 Temperature Testing

Next, we evaluate the quality of the 2-meter air temperature and 2 meter dew point temperature simulation over all five of the simulation domains. Like precipitation, our analysis begins with the examining the seasonal variation in the mean temperatures, as shown in figure **??**. These values are calculated by finding every day in the corresponding dataset that belongs to the given month, finding the mean, maximum and minimum of the corresponding day and then taking the mean of all the days in the dataset. All five of these domains show very strong agreement between the simulations and the observations, with a nearly indistinguishable average in mean, minimum and maximum for the air temperature. The results for the dew point temperature values are weaker than air temperature simulations, with clear evidence of reduced simulated variability compared to observations. Still the average values are consistently well represented. Unsurprisingly, the spread between minimum and

maximum values is considerably smaller for the dew point temperature than it is for the air temperature. The dew point values are also consistently smaller than the air temperature values, which is expected given the definition of dew point temperature, but important to validate as it is not a requirement directly imposed by EGGS-WG. Interestingly, the dew point values seen over the New Guinea is very tightly grouped showing only a small difference between minimum and maximum values.

Following the approach used to compare the spatial correlations for precipitation values in figure **??**, the set of correlation coefficients for every station pair of temperatures is shown in figure **??**. Both the 2 m air temperature and the 2 m dew point temperature are ~~analyzed~~analysed, as well as the cross correlation between these two variables. The cross correlation between air temperature and dew point temperature is important to examine, as this capability should be indirectly built into EGGS-WG. As the temporal resolution of the simulated data is hourly, instead of daily, a slightly different approach is needed than for precipitation. The most simplistic approach to account for these cycles is to separate the data by both the month of the year and the time of day. This does result in 24 times as many grid point correlations as in figure **??**, but should not have any other significant impacts on our analysis.

The temperature correlation results in figure **??** show a relatively large spread of values compared to the precipitation equivalent in figure **??**. Focusing on the air temperature, we see the highest point density is tightly clustered around the 1:1 line, with minor variance either side of the line. This indicates that the spatial dynamics of air temperature are being well captured by the model in general. While there are some outlying values, the difference in magnitude of the correlation coefficients is relatively small. These are all relatively tightly clustered with only the New Guinea region showing some outliers, although this may be at least partially due to having a much larger range of observed correlation values than the other regions. Similar, but weaker, correlations are seen in the dew point temperature difference. One key difference with the dew point temperature data is that the regions of highest density are no longer along the 1:1 line, instead they show a slight deviation indicating a weaker correlation in the simulation than the observations. These differences are minor enough that they should not have a large impact on model accuracy. Interestingly the New Guinea region shows much weaker simulation than all the other cases with the majority of the observed correlations very close to 1 while the simulated correlations are closer to 0.8. We suspect that this is due to statistical issues associated with the minimal range of dew point temperature values over this region as evidenced in figure **??**.

The cross correlation between air temperature and dew point temperature shows more substantial differences between simulations and observations. This is somewhat to be expected as it is combining the variance associated with both of the temperatures. The cross correlations are well represented over the Canterbury region but show clear issues over the other regions. West Coast and Taupo show the bulk of simulations underestimate the correlations, but the effect is relatively minor. The Australia region shows a very wide range of correlations both observed and simulated that generally follows the expected 1 to 1 line. The New Guinea region is very complex with two clear peaks, one well represented and one not. Issues over this region could be related to the earlier identified dew point temperature issues, although further investigation would be needed for certainty. The cause of this regional difference is hard to determine, but may be related to the lower rainfall occurrence over the relatively well represented Canterbury and Australia regions. The highest density of grid point correlations fall relatively close to the 1:1 line, however, with a larger separation than seen in either the air temperature or dew point temperature data. This will

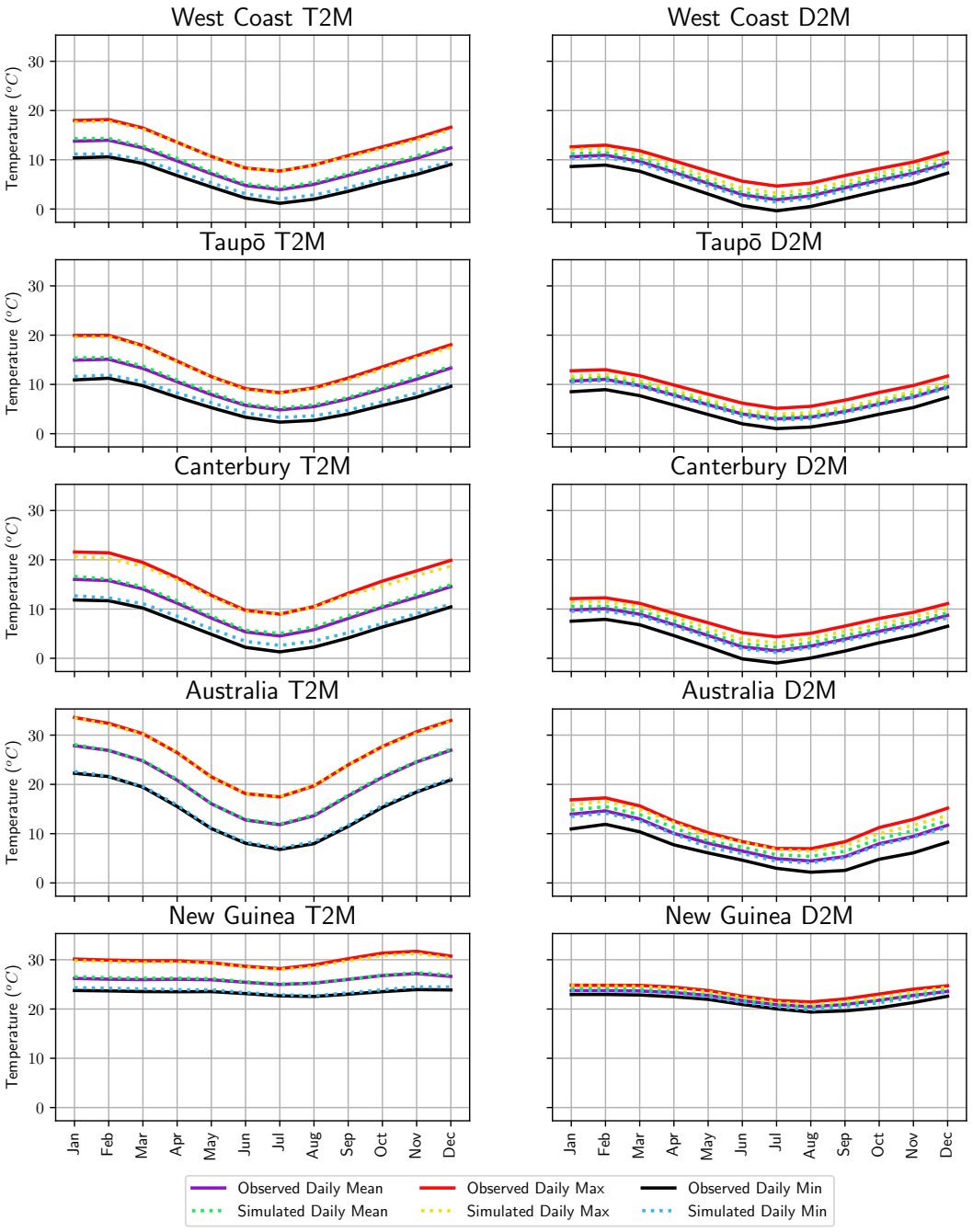


Figure 6. The seasonal cycle of the 2 meter air temperature (T2M) and the 2 meter dew point temperature (D2M) for both the observations and simulations over all five of the domains. Included for each plot is the daily mean, minimum and maximum value for each month of the year. Specifically, the maximum and minimum values here are calculated by finding the daily regionwide minimum and maximum values and then averaging this value over the month.

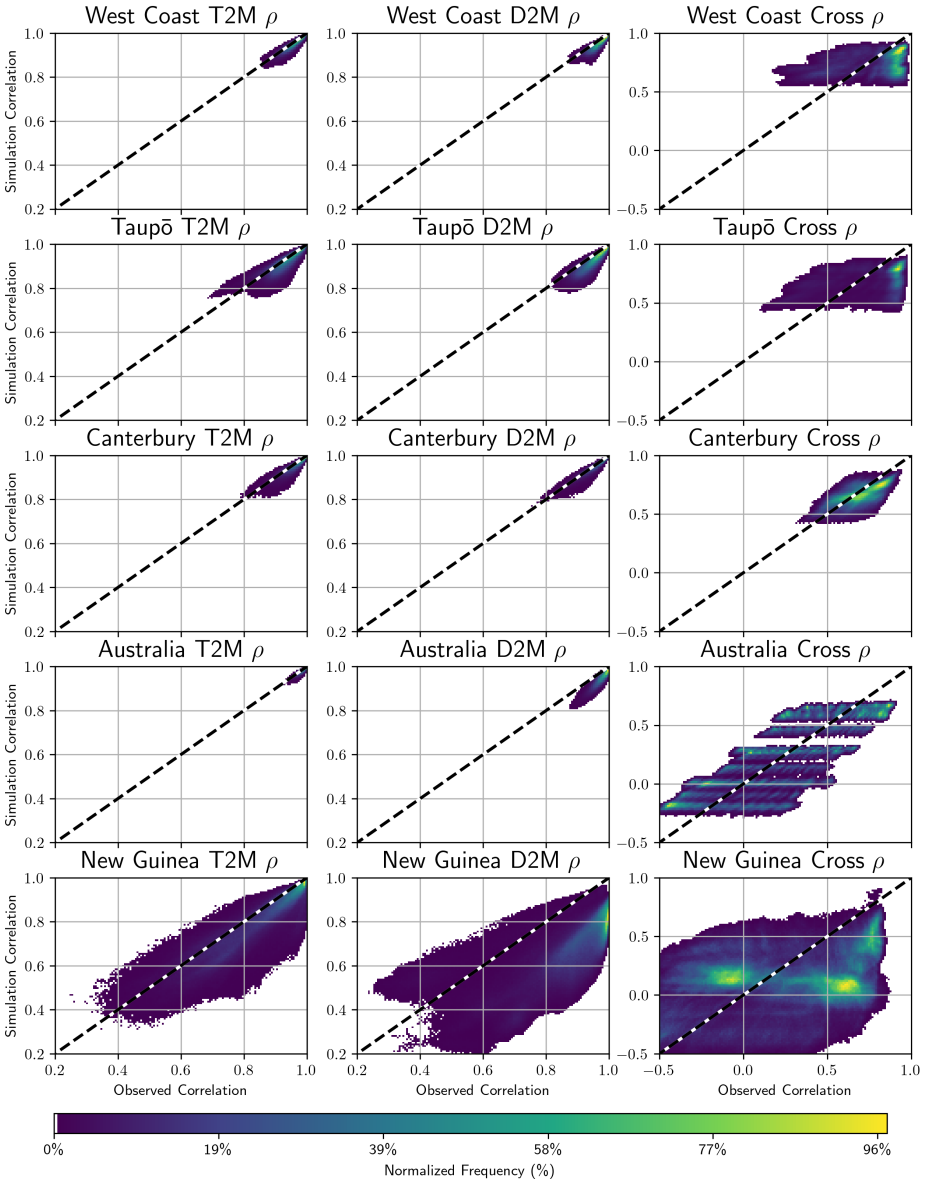


Figure 7. The Pearson correlation coefficient between every set of station pairs for both the simulation and observational data. Values are calculated for both the 2 m air temperature and 2 m dew point temperature data, as well as the correlation between these two variables. The color of the points indicates the density of station pairs at a given correlation, ranging from purple for a low number of points to green for a high number of points. The included dashed line indicates the 1:1 correlation line that would be the theoretical ideal for a simulation.

have implications for the evaluating some simulation outputs, but the correlations still appear close enough to the observed correlations for the majority of purposes.

As the temperature simulations are hourly, it also needs to be confirmed that the simulation is accurately capturing diurnal variability. The representation of the diurnal cycles of both air temperature and dew point temperature is examined in figure **??**. Over all five of the regions, the air temperature values show very strong agreement between observations and simulations. The seasonality is very well captured, which is expected given the earlier results in figure **??**. The dew point temperature is also well represented, but does show some minor differences between the observations and simulations. Generally, the representation is relatively consistent over all five of the domains. While the New Guinea region shows a different annual variability pattern than the other regions, this is unsurprising due to its location in the tropics.

While figures **??** and **??** establish that the average temperature values are well simulated by the model, they do not focus on the representation of the wider distribution. To examine the behavior over the entire distribution, a quantile-quantile plot is included in figure **??**. Note, this value is calculated separately for each grid-point, month of data and time of day. This means that for example, a single dot will show the 50^{th} percentile of all values at 1 pm for a single month, effective removing biases associated with seasonal cycles from the analysis. Figure **??** only shows values for April specifically as seasonal cycles make the plot very difficult to interpret. The patterns seen here are highly representative of all months, with two extra months (January and August) included separately as examples in the supporting information. For both the air temperature and dew point temperature, the simulation very accurately recreates the observational temperature values over all five of the investigated domains. There is a slightly better clustering to the 1:1 line in the air temperature than the dew point temperature, however the differences are relatively minor. Unlike the precipitation results shown in figure **??**, the extreme values of the distribution do not appear to show any significant deviation from the 1:1 line. This suggests EGGS-WG would be more suitable for studying temperature extremes than it is for studying precipitation extremes. The relatively small range of dew point values over the New Guinea region also suggest the issues identified for this region in figure **??**, are more statistical artifact than significant bias.

To further explore any potential temperature biases, we include figures showing the distribution over different timescales, spatial sizes and separately for temperature extremes in the supporting information (figures S 13 - 42). The mean T2M and D2M values appear very well represented by the model over all time scales for all five of the different regions, both over all individual gridcells and in the regionwide mean values. Interestingly, the maximum values t2m are generally well represented, but there appear to be some issues with replicating the coldest temperatures produced by ERA5 and with the yearly d2m maximums, particularly over the Australian region. Given the climates of the specific regions we have examined, the ERA5 values appear to be unrealistically low and may warrant further investigation. However, since the EGGS-WG is purely based on the ERA5 statistical profile, it does highlight a minor failure in the simulation process. The majority of these results do reinforce that the model is in good alignment with the ERA5 output for all but the coldest temperatures.

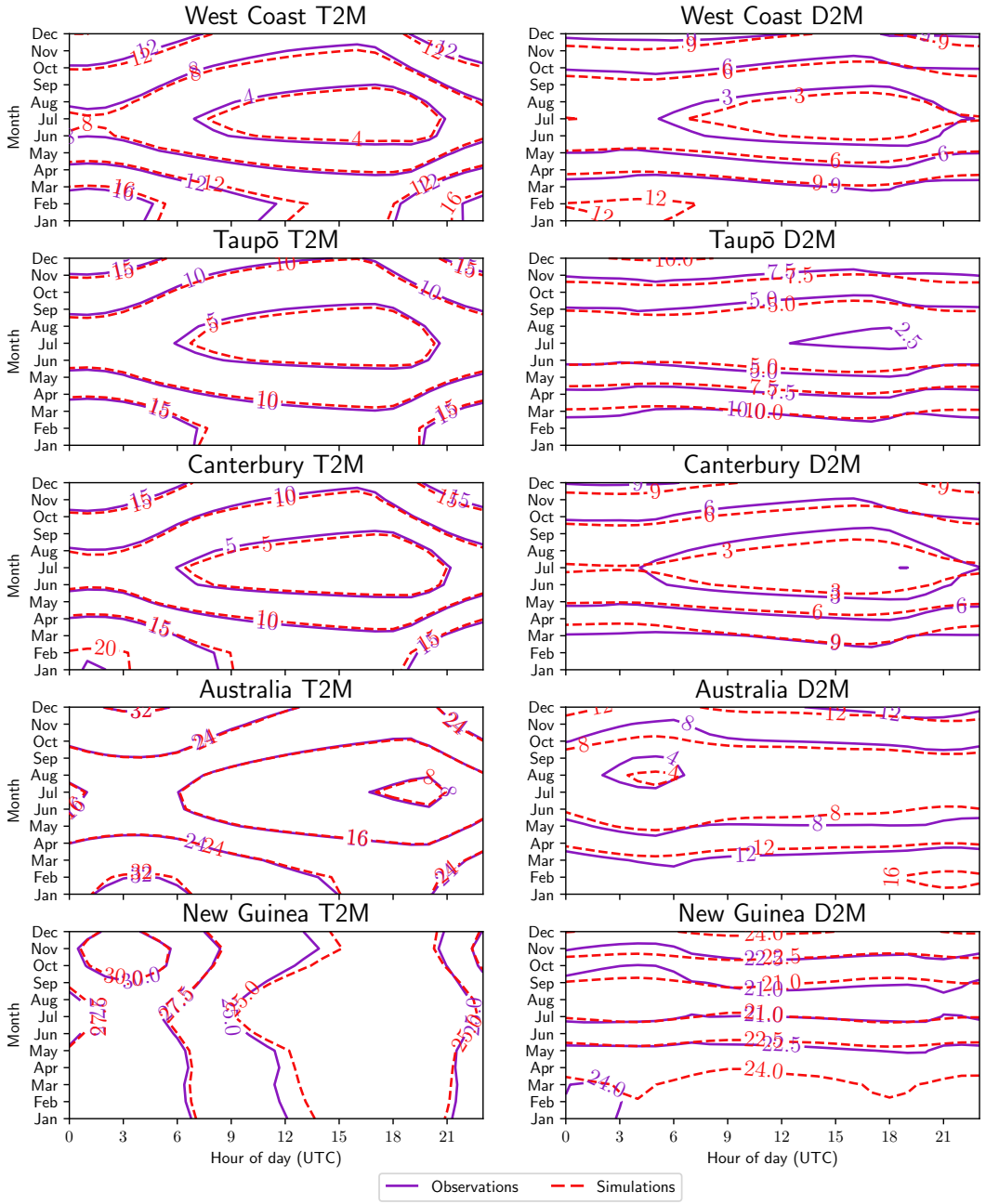


Figure 8. Contour plot of the diurnal variation of 2 m air temperature and dew point temperature for each month in both the simulations and observational data. Temperature values are taken as domain wide average values. Note that time on the x-axis is given in UTC and NZST is +12:00.

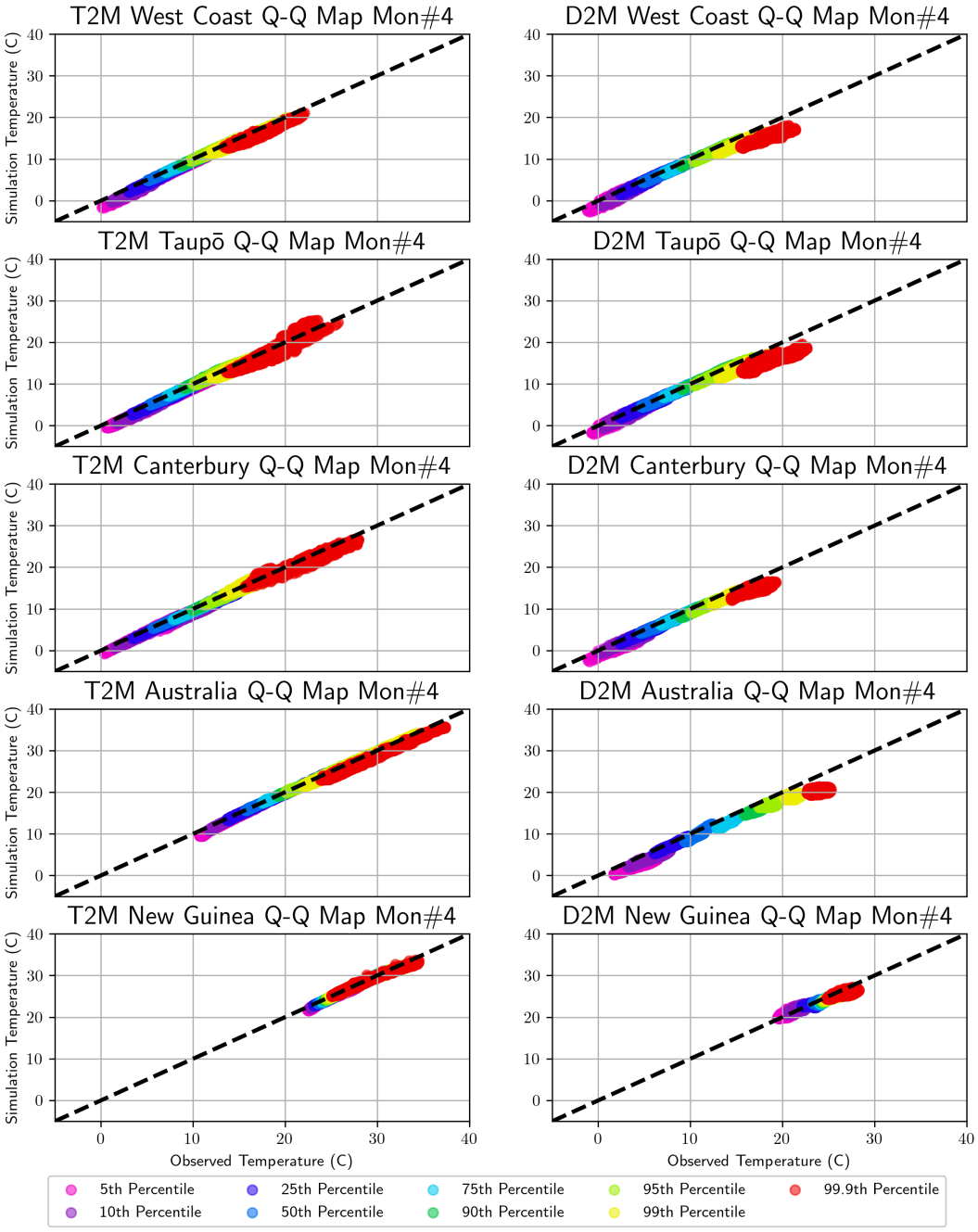


Figure 9. Quantile-quantile comparison of the 2 m air temperature and dew point temperatures for April. Values are calculated independently for each grid point in the domain and for each hour of the day. Each dot represents an individual grid point at a given hour for a corresponding percentile. The percentiles included in this plot are the 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, 95^{th} and 99^{th} percentiles as indicated by the color of the dots. The dashed line indicates the 1:1 correlation line.

4.4 Model Ensemble Variability

The earlier analysis, while comprehensive, is focused on a single model realization. One of the major benefits of stochastic weather generators is their ability to produce a range of simulations with near identical long-term statistical properties but generated from a wide range of realistic time series. This has significant benefits for understanding variability and presenting information to end users that can be interpreted with a storyline approach. To create this ensemble, each of the different simulations are separately initiated with the process described in section **??**, which results in a completely different starting point for each simulation.

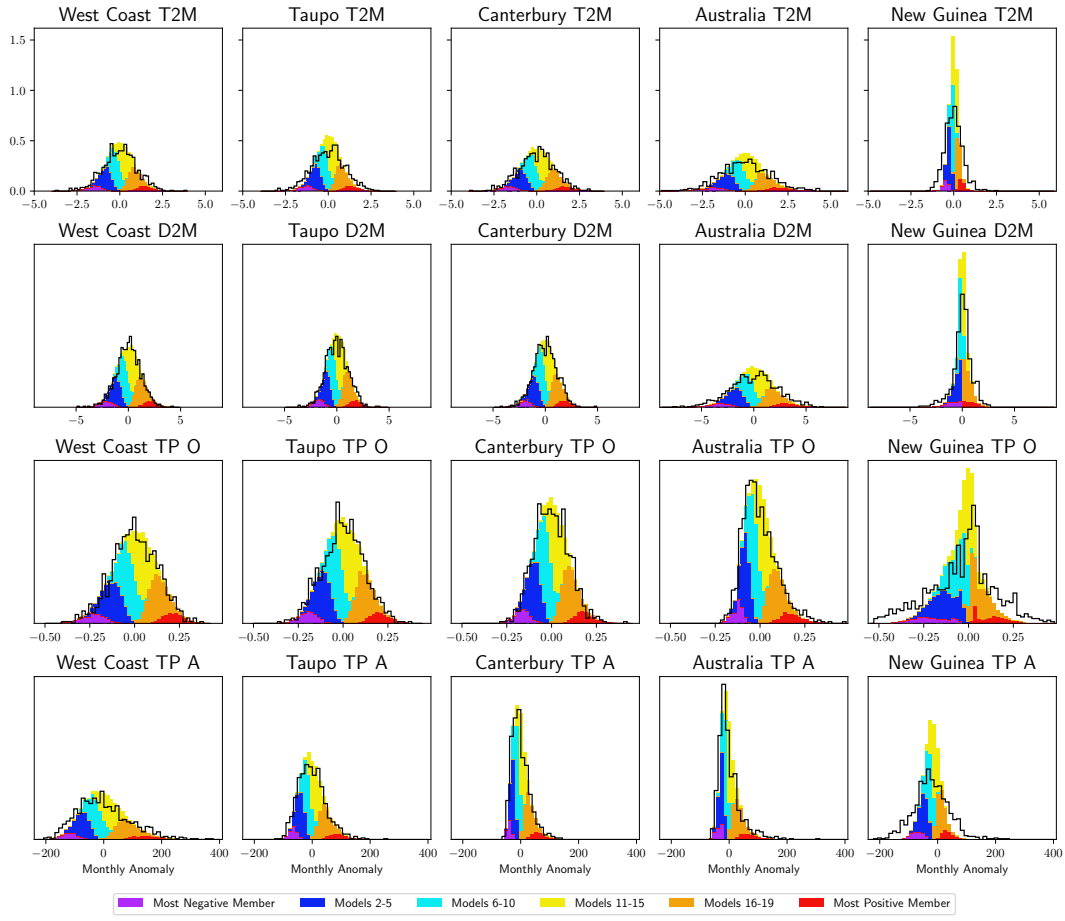


Figure 10. Monthly anomalies of rainfall occurrence, rainfall amount, 2 m air temperature and 2 m dew point temperature for 20 different simulation realizations. The coloring of the histogram indicates the ranking of ensemble members while the black line indicates the equivalent distribution from the ERA5 Land data. Each run is ranked independently for each month.

Figure **??** shows is the ranked anomalies of our 20 run ensemble for 2m air temperature, 2m dewpoint temperature, rainfall occurrence rate and accumulated rainfall. Effectively each month of our 20 model simulation is considered separately, and the

models are split and ranked based on the size of their monthly anomaly. The normalized 20 model ensemble distribution is then compared directly to the equivalent values calculated from the ERA 5 Land dataset. This figure is included to test a few different features of the simulation; the quality of the data when aggregated at a monthly level, the range of outcomes produced across an ensemble for a given month and how any biases identified in these distributions differs over the different regions.

Examining the 2m air temperature in figure **??**, we see a consistent Gaussian of similar width produced over all of the regions, strongly inline with the observed distribution, with the exception of the New Guinea region which shows a much more narrow Gaussian. A good spread between ensemble members is seen in the other four regions with evidence of notable variation even in the 1^{st} and 20^{th} ranked ensemble members. The results for New Guinea are substantially different from the other regions with a significantly narrower distribution shown in the simulations relative to the observations. Amongst the ensemble over New Guinea there is also evidence of the most extreme ensemble members showing anomalies close to zero. Very similar biases are seen for the dew point temperature, but with a slightly improved quality of simulation of New Guinea. For the precipitation occurrence and accumulation, we see strong agreement between simulations and observations in the first four regions including for some distributions that are clearly non-Gaussian in shape. For the New Guinea region we see some distinct behavior that disagrees with the observations, similar to the temperatures this is mostly in the form of a reduced variability. Interestingly, in the precipitation occurrence there are cases with the most extreme ensemble members showing the monthly mean value. These are most likely due to the large variations due to the monsoon signals, especially during times of close to 100 percent precipitation occurrence. Overall, the results presented show that the model does a good job replicating the observations and the ensemble is showing a reasonable degree of variability. The only exception is the New Guinea region which still shows a reasonable simulation but clear biases.

In addition to figure **??**, we include two figures in the supporting material one which shows the monthly anomalies in the ensemble as time series (figure S43) and one that shows the distribution of Pearson correlation coefficients between each of these different ensemble members (figure S44). Specifically, this is the correlation coefficient between time series of monthly mean 2 m air temperature, 2 m dew point temperature, precipitation occurrence and cumulative precipitation between all ensemble members. From these correlation coefficients we can see that the different ensemble members are effectively uncorrelated from one another (with the exception of a minor correlation in the New Guinea data). Combining this information with the results shown in figure **??** and **??**, which demonstrate that the individual model runs are successfully reproducing the variability within the observations, we can conclude that the individual ensemble runs are able to provide sufficiently variable simulations that are unrelated to each other. As discussed earlier there are some issues at the most extreme portions of the distribution but in general the variability seen between the ensemble members is good.

4.5 Extra Simulation Regions

Given the problems established with the simulation of the ~~Papua~~ New Guinea region, we expanded the analysis over this area to include several subregions. Included in the supporting information is figure S45 that shows four subregions that we label PNG 1, 2, 3, and 4. Also included are versions of several of the main figures from this paper restrained to these new simulation domains, Figures (S46-49), so that we can determine if any simulation issues are related to the size of the domain or the

region. Immediately clear is that the occurrence correlation issue is not present in any of the subdomains~~, meaning it must be a result of the size of the domain (or more specifically, the variation of the domain)~~. This is most likely due to the range of conditions that can occur over the domain, with evidence from later analysis of large regions suggesting a direct relationship to the climate over the region. Interestingly, the issues identified with streak length are prevalent in all sub-regions, so they ~~are likely due to a~~ must be due to an underlying limitation from using such a simple model of precipitation occurrence. ~~We~~ The poor representation of streaks exactly one day long would be directly resolved by using a depth-2 markov chain however the impact on longer streaks is unclear and we aim to investigate this further~~and potentially improve rainfall performance over this region in the future~~. Finally, the temperature correlation plots show some improvement, but issues clearly remain in the T2M D2M cross correlation. One possibility is that the relatively small local variations in D2M that are reasonably well captured over the smaller domains lead to spurious relationships in the correlation coefficeints.

To further explore these issues, we also include the results from 5 simulations over North America and 5 simulations from the Indian Subcontinent (figures included in the supporting information as S50 - 59). Both of these regions have four 225 gridcell simulations and one 625-grid-cell simulation (to match the New Guinea region). The precipitation correlation results over the North American regions are exceptional, while for the subcontinent regions, performance is good but similar to the New Guinea region, showing a wide spread. Similar results are seen in the streak values (exceptional NA, average IND), supporting the claim that the New Guinea biases are ~~possibly~~ directly related to the monsoon climate. Both regions show better representation of precipitation extremes than the other regions of our paper, which is a positive sign that our identified extreme precipitation issues are not universal. Finally, the temperature correlations are also examined. The values over the North American domains show excellent agreement with observations, including both the cross correlation and for the larger region. For the subcontinent region, we see great temperature correlations over two of the regions, and then the remaining three regions show similar behavior to the New Guinea regions. Interestingly, all three of these regions show only minor variation in D2M and are impacted by the monsoon, with two of them strongly impacted. Overall, the two bigger domains show better agreement with observations than the New Guinea region, reinforcing that the particular climate of New Guinea is what causes the model issues. This argument is supported by the monsoon impacted Indian Subcontinent regions showing biases similar to the New Guinea region.

5 Conclusions

This paper covers the development of a new open source freely available stochastic weather generator, EGGS-WG. The structure and design of this model is based on a well established research framework covering Richardson style generators (**??**) which has guided the design of many of the most prominent existing stochastic weather generators (**????**). However, many of these weather generators are lacking features that have been established in research following these initial models (**??**). Modern features of EGGS-WG that are missing from older models include a gridded multisite approach, hourly temperature simulation, simulation of dew point temperature and relative humidity and a global terrestrial simulation domain. In particular, the movement to the ability to simulate any region of the planet is a major development compared to existing models. By

designing this model for portability, low resource requirements and ease of use, it is our aim to provide software that can useful for as wide a group of end-users as possible.

Comprehensive validation of the EGGS-WG output is performed throughout this paper, focusing on three distinct regions over New Zealand, one in Australia and one in New Guinea. There are four key variables produced by EGGS-WG that are examined to evaluate performance; precipitation occurrence, precipitation amount, 2 m air temperature and 2 m dew point temperature. The mean values for all of these variables are investigated and show a strong agreement between the reanalysis and simulations. Further tests show that the seasonal cycle, diurnal cycles (for temperature only) and spread of the distribution of these variables is also well captured by the simulations. There is an identified reduction in quality in the simulation of high-end precipitation extremes (~~above~~ 99^{th} percentile and above), although the rarity of these points limits the impact on the mean state of the simulation. This does however mean that caution should be used if the model is being used expressly for the purpose of extreme simulation. The ability of EGGS-WG to recreate observed spatial relationships is also tested and shows generally strong results, except for the cross correlation between air temperature and dew point temperature. The simulation of this relationship is still reasonable but does show consistent underestimation of the observed relationship. Additionally, the model is shown to produce variable time series that can produce a wide range of synthetic story lines that accurately capture the aggregate statistical properties of the observations.

Currently, there are a few major limitations to this model. First and most important is that the model is based entirely on the ERA5-Land data and as such will share any biases present in the ERA5 data. Given all of our comparisons are made to ERA5 data as observations, this also a limiting factor of our model evaluation. The most commonly discussed issue with ERA5 is a limited capability to represent the most extreme weather events, particularly underestimating precipitation extremes. These issues may be improved with the upcoming release of ERA6 which we hope to integrate into our model quickly. The model simulation has also been shown to have a flawed representation of both precipitation extremes and minimum temperatures. Although the model tends to bias towards producing too much extreme precipitation and the ERA5 produces too little, so the exact impact is difficult to determine. There are other potential limitations including accurate representation of domain wide weather events, the impact of a low depth Markov chain on accurately capturing streaks (particularly in ver low or high precipitation occurence regions) and lack of relationship to global teleconnections such as ENSO, which are not evaluated here but may be a focus of future work. These limitations lead us to make the following recommendations for using the model:

- General best practice with this model is to work on regions below 1000 grid cells (it will work, but it gets increasingly computationally inefficient), generate ensembles where possible, and use the stationary mode unless you have a reason not to. The model output should be suitable for the majority of tasks that SWG output is used for. If you have questions or concerns about a given application, the corresponding author can be emailed directly (Alex.schuddeboom@canterbury.ac.nz) or a post can be made on the github issues page (https://github.com/alexschuddeboom/EGGS-WG/issues).

- Due to identified issues with precipitation extremes (particularly those greater than the 99^{th} percentile), extra caution should be used if this is the primary focus of your analysis. The first thing to do is acknowledge with any of your results that the precipitation tends to bias towards larger precipitation values in your results. We also recommend simulating

multiple different ensemble members and examining the range in extremes between these values. There is a lot of extreme research focused on values that are not quite as far along the distribution (ie 95^{th} percentile) and these issues do not impact those values. At the 99^{th} and 99.9^{th} percentile, the bias is highly variable between different regions and larger as a percent of rainall for lower rainfall regions, meaning the researchers who should be particularly cautious are those looking at absolute maximum rainfalls for work such as flood analysis. For some regions examined we do not see this bias, if you would like a region validated, the authors are open to collaboration.

- If examining temperatures, the simulation performs very well in the majority of cases. The two clear places with potential issues identified are dew point temperature variability and the extreme minimum observed temperatures. If either of these is the primary focus of your work, this weather generator may not be suitable. We do have plans to address these going forward.

- There is clear evidence of weaker model performance over regions that are strongly impacted by the Monsoon. Usage of the model within these regions should be done cautiously, with acknowledgement of the limitations. Additionally, if the model us used on a region with no existing comparisons, short validation tests against both ERA5-Land and independent datasets are recommended. This include regions with no Monsoon activity as there is the possibility of other regional biases.

- Table S1 in the supporting information shows some times required to run existing models on a laptop and server. For a laptop simulation of a smaller domain (less than 200 gridcells) should range from 3-8 hours. The code has been improved since these benchmarks were recorded, so they can be considered an upper bound. On the current code a server simulation of a 1.5° domain takes less than an hour and 2.5° less than two hours. As the domain size increases, the computational costs increase (in our experience, somewhere between order N and N^2 but exact characterization is complex); the main concern is RAM, although our benchmarks do show that a laptop with only 16gb of ram can run a smaller domain. It is our recommendation that if you are simulating larger domains, ensure your system has enough ram.

There are also many different future updates being considered for the EGGS-WG model. Most prominent amongst these is adding a climate change forcing component to the model that would allow for direct integration of CMIP6 model data to impact future climate. The second priority is an expansion of the model to include support for other climate variables, most importantly wind vectors, incoming radiation and surface pressure. We have developed an internal version of the model that supports both of these feature, but substantial work on data quality, portability, data accessibility and ease of use is required for them to be added to a public release. There are more speculative features being considered, including integration with existing crop growth models (**??**) or streamflow models (**??**), comparison with up and coming GAN based stochastic weather generators (**?**) and exploration of alternative precipitation schemes to drive an hourly precipitation model. There is also the possibility of future research based on the direct output from the EGGS-WG model. Future work may also move beyond ERA5-Land and rework the model to also work with other modern datasets.

Code and data availability. The model produced is open source and publicly accessible on GitHub under a GPL-3.0 license (https://github.com/alexschuddeboom/EGGS-WG). There is a specific version of record for this paper, V 1.2.1, that is also accessible on Zenodo at https://doi.org/10.5281/zenodo.19228384 (**?**). To simplify the ERA5-Land data access process, included with our model is a python script which can be used to download ERA5-Land data directly through their API. Also included on GitHub is a technical manual that briefly describes the operational procedure for the model.

Author contributions. Alex Schuddeboom: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing - Original Draft, Writing - Review & Editing; Christian Zammit: Conceptualization, Funding acquisition, Project administration, Supervision, Writing - Review & Editing; David Plew: Conceptualization, Funding acquisition, Supervision, Writing - Review & Editing; Piet Verburg: Conceptualization, Funding acquisition, Supervision, Writing - Review & Editing; Aidin Jabbari: Supervision, Writing - Review & Editing

Competing interests. The authors declare that they have no conflict of interest.

Acknowledgements. The authors would like to acknowledge the feedback of all reviewers that have provided input, which has refined and improved the paper. The authors would like to thank the European centre for Medium-range Weather Forecasting for the production of the ERA5-Land dataset, without which this model would not be possible. The ERA5-Land data is accessible through the Copernicus Climate Change Service Climate Data Store **?**. Additionally, we would like to thank the Copernicus Climate Change Service Climate Data Store for making the data widely available and publicly accessible. Funding for this project was provided through both the New Zealand Ministry of Business, Innovation and Employment (MBIE) via the 2022 Endeavour Fund - Smart Ideas (contract C01X2205) and by the Aotearoa New Zealand Tāwhia te Mana Research Fellowships, administered by the Royal Society Te Apārangi (contract MTP-UOC2401).