Report on SLMACC Climate-Shock Vulnerability: Infinite Improbability Drive and Stochastic Weather Generator

Bryn Gibson, Alex Schudeboom, Greg Bodeker, Annabel Yip

3 February 2021

Table of contents

Introduction

The purpose of this SLMACC (Sustainable Land Management and Climate Change) project is to analyse the impacts of climate shock-events to identify farm system vulnerability, sources of resilience, and options to mitigate risk and promote proactive adaptation.

In support of the overall project, Bodeker Scientific was contracted by Water Strategies Ltd to deliver:

1. An **Infinite Improbability Drive (IID)** that, for a given storyline of temperature, precipitation conditions and irrigation restrictions, quantitatively defines the likelihood of that storyline occurring.
2. A **Stochastic Weather Generator (SWG)** that, for a given storyline, can generate time series of daily:
    a. Maximum and minimum temperature (T_{max} and T_{min}).
    b. Broadband radiation.
    c. Precipitation
    d. Potential Evapotranspiration (PET)

In both cases the storylines are provided by Komanawa Solutions Ltd (KSL).

The purpose of this document is to provide a functional description of these two components of the SLMACC project processing chain. Sub-sections hereafter define some of the key concepts underlying both the IID and SWG.

Storyline and state definition

A storyline is defined as a chronological sequence of months classified in three primary states:

1. **Temperature:** hot (H), cold (C) or average (AT). Hot months were defined as any November, December, January, February or March with 10 or more days experiencing T_{max} greater than or equal to 25°C. Cold months were defined as any May, June, July, August, or September with 10 or more days where the 3-day running mean daily average temperature (mean of T_{max} and T_{min}) is less than or equal to 7°C. Months classified as neither hot nor cold were labelled average (AT).
2. **Precipitation:** dry (D), wet (W) or average (AP). Dry months were defined as any month other than June or July when the soil moisture anomaly was less than -20 mm for 10 days or more. Wet months were defined as any May, June, July, August or September when the number of days with 0.1 mm or more of rain was greater than or equal to *N*, where *N* depends on month, viz. May (*N*=14), June and July (*N*=11), August and September (*N*=13). Months classified as neither wet nor dry were labelled average (AP). Note that the W and D classifications are not mutually exclusive. For example, there could be a May that meets the criteria for both W and D; this occurred in less than 0.5% of the classifications so is not expected to significantly bias the classification. In such cases, the month is classified as dry (D) because dry months have a larger impact on the pasture growth model.
3. **Irrigation restriction:** a value between 0 (no restriction) and 1 (full restriction).

The 'state' for any month is then defined by values for each of these states, e.g. 'H,D,0.2' would denote a hot, dry month with a 20% irrigation restriction.

Virtual Climate Station Network (VCSN) data

The VCSN data set provided by NIWA covering the period 1 January 1972 to 24 March 2020 constituted the primary source of weather/climate data for these analyses. The VCSN data set for the New Zealand land surface is constructed on a regular 0.05° x 0.05° grid from spatially inhomogeneous and temporally discontinuous quality-controlled weather station data (Tait et al., 2005). The values estimated on the 0.05° x 0.05° grid are based on thin plate smoothing spline interpolation using a spatial interpolation model as described in Tait (2008). For the purpose of this study, the following daily data were used from the VCSN data set: precipitation, T_{max} and T_{min}, surface downwelling shortwave radiation (rsds) and potential evapotranspiration (PET).

When used to derive quantiles for calculating H/C/AT and W/D/AP states for the weather@home data (see below) and the basis for the SWG, it is important that the VCSN data are 'normalized' to the present. This was done by fitting the following equation to the time series of each target climate variable at each grid node of the VCSN data set:

$$V(y) = a_0 + a_1 \times T'_{SH-land}(y) + \sum_{i=1}^{3}(a_{2n} \times \sin(2i\pi y) + a_{2n+1} \times \cos(2i\pi y))$$

where *V* is the variable of interest, $T'_{SH\text{-}land}(y)$ is a time series of the southern hemisphere land temperature anomaly with respect to pre-industrial, *y* is the decimal year, and the a_n are fit parameters determined through linear least squares regression. The Fourier expansion ensures that the derivation of the trend, quantified by a_1, is not affected by the seasonal cycle which is fitted by the Fourier terms. A normalization to 2019 was then derived from the $\boldsymbol{a_1}$ coefficient and applied to all VCSN data prior to 2019. This normalization was applied to all relevant VCSN variables except for precipitation - only precipitation values not equal to 0 were adjusted and any adjusted below zero were made positive.

weather@home simulations

To calculate robust state likelihood and transition probabilities, the IID makes extensive use of weather@home simulations (Black et al., 2016). weather@home simulations are generated in large initial-conditions ensembles that each manifest as the same climate but different weather. Each ensemble is provided as a 'batch'. For the purposes of this analysis, 1400 simulations from weather@home batch 793 were used. Each simulation in this batch covers the 20-month period December 2016 to July 2018 and, as such, is representative of the current climate. To avoid different sample sizes in different months, and to avoid the earlier months that may be prone to model spin-up irregularities, we used the 1-year period from July 2017 to June 2018 for the analysis. 64 of the 1400 were excluded from the analyses since they had one or more NaN values (denoting missing data).

Because the model underlying the weather@home simulations uses a 360-day calendar (30 days in each month), the first 30 days in each month in a weather@home simulation were mapped to the corresponding days in reality. For February, the additional 2 days in the weather@home simulations were mapped to the average of 28 February and 1 March.

To determine the soil moisture anomalies required for classifying precipitation states, rather than using the soil moisture values from weather@home which were deemed unreliable, daily soil moisture values were calculated by applying the Penman-Monteith equation taking weather@home fields of short-wave radiation, precipitation, air temperature, relative humidity, wind speed, mean sea level pressure, and elevation as input.

The weather@home data are very likely to be biased against the VCSN data for two reasons:

1. They are output from model simulations that will have an imperfect representation of the dynamical and thermodynamical processes affecting the climate over New Zealand.
2. The model simulations are performed at fairly coarse resolution (0.44° x 0.44°) that will not resolve topography and land surface features at the same resolution as the VCSN data set which results from gridding station measurement series.

A quantile-mapping method was used to correct for any such biases. Consider a calendar day *D*. Data are then considered for day *D* and 5 days either side of *D*. For 48 years of VCSN data there are expected to then be 48 x 11 = 528 data, each classified as H/C/AT and W/D/AP. The 528 values are then sorted and the fraction (percentile) of values falling within each classification is determined. These percentile statistics were provided by KSL.

As an example consider temperature for 1 January: the percentile for ‘hot’ is 68.939%, indicating that 33 of the 48 T_{max} values on 1 January, and 5 days either side of 1 January, in the VCSN data were below 25°C, i.e. were average (Januaries cannot be cold as per the classification schema), while 31.061% were classified as ‘hot’. To then classify all of the 1 January weather@home days, the T_{max} values for 1 January across all ensemble members are sorted and the hottest 31.061% are labelled as ‘hot’ days. A similar quantile mapping approach is used to ensure that a similar fraction of the weather@home days are labelled W/D/AP as in the VCSN data set.

Once every day in the ensemble of weather@home simulations was classified as H/C/AT and W/D/AP, each month in the weather@home simulation can be similarly classified using the definitions provided above. It should be noted, however, that this can result in a different fraction of months being labelled, e.g., ‘dry’ in the weather@home ensemble than in the VCSN data set. Table 1 lists the differences in percentiles of months with that classification between the weather@home ensemble and the VCSN data set. There is clearly some resultant bias in the fraction of May months classified as H/C between the VCSN and weather@home. These biases most likely arise from the day-to-day variability in weather@home temperature and precipitation records differing slightly from that in VCSN, as well as the difference in the spatial scale of the data sets, as a single weather@home grid cell contains roughly 100 VCSN grid cells.

	D	***AP***	***W***	***H***	***AT***	***C***
Jan	0.96%	-0.96%		-8.26%	8.26%	
Feb	7.31%	-7.31%		-10.74%	10.74%	

Mar	3.58%	-3.58%		-9.93%	9.93%	
Apr	0.54%	-0.54%			0.00%	
May	-2.93%	3.85%	-0.92%		-17.68%	17.68%
Jun		1.19%	-1.19%		0.41%	-0.41%
Jul		4.82%	-4.82%		-2.02%	2.02%
Aug	-6.52%	5.70%	0.82%		-7.83%	7.83%
Sep	-5.31%	5.65%	-0.34%		-0.42%	0.42%
Oct	-0.15%	0.15%			0.00%	
Nov	3.80%	-3.80%		-5.49%	5.49%	
Dec	2.08%	-2.08%		-1.38%	1.38%	

Table 1: The differences in percentiles of each month in each classification between the VCSN data set and the weather@home data set. Negative values indicate where a classification is less likely to occur in VCSN than in the weather@home data set. A zero value indicates that both data sets have the same fraction of months with that classification. Greyed cells show where that classification, by definition (see above), is not possible and is therefore 0% by definition.

To provide a partial validation of the bias correction approach applied to the weather@home data as described above, a second method was implemented where the threshold values applied to classify the weather@home days as H/C/AT and W/D/AP were adjusted to ensure that the same fraction of individual weather@home days have a given classification as in the VCSN data set. This was done by calculating the average daily percentile value for each class across all possible months and finding the corresponding threshold. For example, the average 'hot' daily percentile was calculated across all possible 'hot' months, i.e. the percentiles in each day in November, December, January, February and March. This average percentile was then used to find the threshold value to classify the weather@home data. These adjusted thresholds to be applied to weather@home classification were:

- Hot: 295.26 K/22.11°C
- Cold: 280.76 K/7.61°C
- Dry: -22.83 mm soil moisture anomaly
- Wet: 2.02 mm/day

The resultant differences in the percentiles of each month in each classification between the VCSN data set and the weather@home data set were significantly worse (not shown) than for the quantile-mapped classification differences (Table 1) suggesting that of the two possible approaches, quantile mapping gave superior bias correction results and so was the approach adopted hereafter.

The Infinite Improbability Drive (IID)

The IID calculates the probability of a given storyline by calculating the probability of the first month's state and then multiplying that probability by the transition probability to the state of every subsequent month in the sequence. In addition, each month in the sequence has an associated irrigation restriction probability. The product of all of the probability terms defines the total storyline probability. Figure 1 shows this schematically.

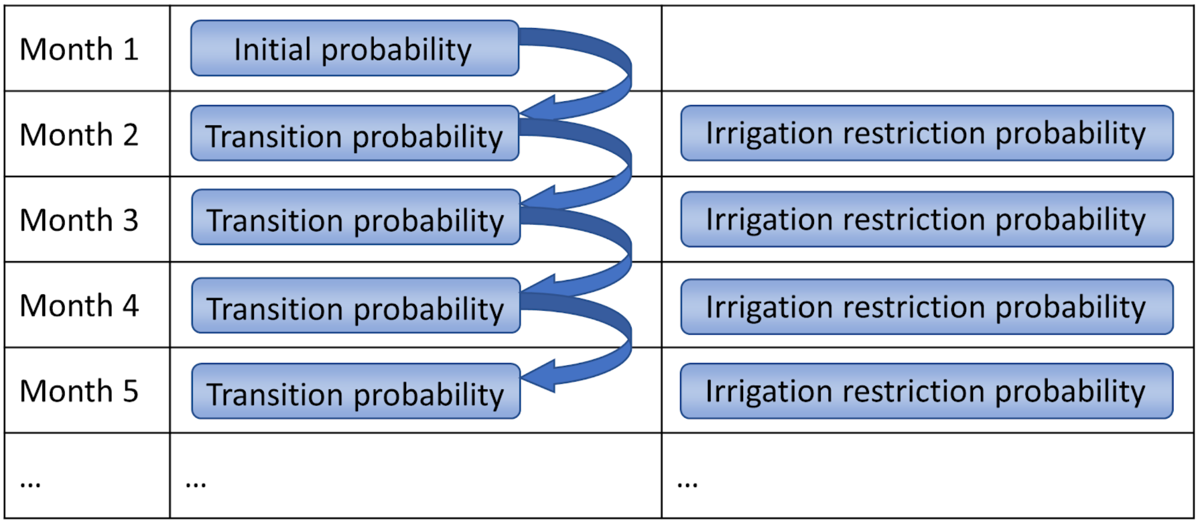

Figure 1: Schematic representation of the IID. Each blue box shows a term. The product of all terms defines the probability of the storyline. Note that because the irrigation probability in month N is a function of the precipitation state in both month *N* and month *N-1*, an irrigation restriction probability is not calculated for the first month.

The large ensemble of weather@home simulations (see above) was used to calculate both the initial state and transition probabilities. Details of the calculations of each of the three terms shown in Figure 1 is provided in the three subsections below.

Initial weather state probability calculation

To calculate the probability of the first month in a storyline being in a certain state, count tables were generated for every calendar month from the classified weather@home monthly mean data. The number of times a calendar month was in a certain state was counted and then divided by the total number of instances across all states.

	AT,AP	AT,D	AT,W	C,AP	C,D	C,W	H,AP	H,D	H,W
Jan	0.277	0.130	0.093	0.131	0.017	0.139	0.087	0.117	0.010
Feb	0.282	0.123	0.101	0.135	0.015	0.131	0.079	0.121	0.013
...	...	...	...	...	...	...	...	...	..

Table 2: An example initial probability table, showing the probability of a calendar month being in a particular weather state.

Weather state transition probability calculation

To calculate the probability of any transition between any pairs of states, transition count tables were generated for every calendar month from the classified weather@home monthly mean data.

Each transition count table shows the number of times in the weather@home data a particular calendar month transitioned from a certain weather state to another weather state in the next month, for all possible combinations. Table 3 shows an example transition count table for January, showing the number of times Januaries in a particular weather state transitioned into Februaries in a certain weather state.

		January			
		AT,AP	AT,D	AT,W	...
February	AT,AP	628	302	215	...
	AT,D	275	105	96	...
	AT,W	222	113	62	...
	...	...	...	...	...

Table 3: An example transition count table for January. January weather states are listed across the columns and the February weather states are listed down the rows. The values show the number of times that particular January state transitioned into that particular February state. In the above example, 628 Januaries which had an average temperature and average precipitation, transitioned into Februaries which had an average temperature and average precipitation. 113 Januaries which had an average temperature and were dry transitioned into Februaries which had an average temperature and were wet.

The transition probabilities were calculated by dividing each value by the sum of its column. For example, in Table 3, the probability of an AT,AP January transitioning into an AT,W February is 222/(628+275+222…). Transition probability tables were generated for each calendar month, showing the transition probabilities of all possible pairs of weather states.

		January			
		AT,AP	AT,D	AT,W	...
February	AT,AP	0.2940	0.3011	0.2994	...
	AT,D	0.1287	0.1047	0.1337	...
	AT,W	0.1040	0.1127	0.0864	...
	...	...	...	...	...

Table 4: An example transition probability table for January. This table shows the probability of a particular January state transitioning into a particular February state. In the above example, the probability of an AT,AP January transitioning into an AT,AP February is 0.2940. The probability of an AT,W January transitioning into an AT,D February is 0.1047.

Irrigation restriction probability calculation

Historical irrigation restriction data were also required for mapping irrigation restriction values to the VCSN data to produce irrigation restriction cumulative density functions (CDFs) for precipitation disaggregation. Irrigation restriction probability is the probability of the monthly mean flow restriction being equal to or greater than the monthly mean restriction value prescribed in a storyline, given this month's precipitation state, and the previous month's precipitation state. Restriction values vary between 0 (no restriction) and 1 (100% restriction). To disaggregate the restriction by precipitation state, months were categorised into either 'not dry' (ND) or 'dry' (D).

To calculate the irrigation restriction probability, the data provided by KSL in the files restriction_record_detrend.csv and event_definition_data.csv were merged, and means of the daily flow restrictions were calculated for each month.

Before the irrigation data was analysed, it was detrended following the same approach used for the VCSN precipitation data. This means that days with irrigation values of 0 were left unchanged, but for all other days the irrigation values were adjusted based on the earlier derived Fourier fit. Additionally, any days with an adjusted value greater than 1 were set to 1.

Using the agreed on definitions for dry, wet and average (in terms of soil moisture) months, each month in the historical irrigation restriction records was classified as either ND or D. For any given month and the previous month, this results in four disaggregation states i.e. (ND,ND), (ND,D), (D,ND), (D,D). For each of these disaggregations, the number of times a restriction value fell within one of 6 regimes was calculated i.e.

1. Irrigation restriction of exactly 0.0.
2. Irrigation restriction between 0.0 and 0.25.
3. Irrigation restriction between 0.25 and 0.5.
4. Irrigation restriction between 0.5 and 0.75.
5. Irrigation restriction between 0.75 and 1.0.
6. Irrigation restriction of exactly 1.0

Normalized probability density functions (PDFs) were then calculated for each month. When subtracted from 1.0, and given an irrigation restriction from a storyline, these CDFs can then be used to calculate the probability of that restriction. For example, using Table 5 below and a prescribed flow restriction of 0.35 in January, linearly interpolating between the bounding values of 0.556 and 0.444 gives a probability of this restriction of 0.5112.

Rest.	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec
0	1	1	1	1	1	1	1	1	1	1	1	1
0	1	1	1	0.846	0.889	Nan	Nan	Nan	1	0.667	0.286	0.7
0.25	0.556	0.714	0.857	0.846	0.444	Nan	Nan	Nan	0	0	0.143	0.4
0.5	0.444	0.571	0.786	0.538	0.222	Nan	Nan	Nan	0	0	0	0.1

0.75	0.111	0.286	0.5	0.231	0.111	Nan	Nan	Nan	0	0	0	0.1
1	0	0	0.071	0	0	Nan	Nan	Nan	0	0	0	0
1	0	0	0	0	0	Nan	Nan	Nan	0	0	0	0

Table 5: Restriction likelihood probability calculation table for D,D months. The first column lists restriction values. Each column lists the likelihood of the flow restriction being equal to or greater than the prescribed restriction value in the first column. NaN values result from there being no data for that month for this disaggregation. Note that for September there was only a single instance of a dry September preceded by a dry August and the flow restriction in that case was 0.

The probabilities calculated using the 4 tables are included as additional terms in the IID to calculate the total probability of occurrence of a particular storyline.

Input and output of the Infinite Improbability Drive

A directory of .csv files comprising a collection of storylines is provided as input to the IID. Each storyline is sorted by year and month, with each month in a storyline having a prescribed weather state (consisting of a temperature state and a precipitation state), and a prescribed restriction value, as shown in the example in Table 6.

year	month	class	temp_class	precip_class	rest
1	1	H,D	H	D	0
1	2	AT, D	AT	D	0
...	...	...	...	...	
1	12	C, AP	C	AP	0.3532
2	1	AT, AP	AT	AP	0.14353
...	...	...	...	...	...

Table 6: An example input table showing each month's weather state (class), disaggregated into its temperature state (temp_class) and precipitation state (precip_class), and restriction value (rest).

The IID calculates the probability of the prescribed combination of events and generates a table showing the probability of each provided storyline, as shown in the example in Table 7. Some stories will have a 0 probability if they contain a transition that was never experienced in the weather@home dataset or if a prescribed irrigation restriction value is unprecedented in the historical record for that combination of current and previous month's D/ND state.

file	prob
v2.csv	0
v1.csv	0
v3.csv	6.12E-13

v5.csv	0
v4.csv	1.57E-06

Table 7: An example of an output from the IID. The ‘file’ column shows the storyline’s file name, and the ‘prob’ column shows the corresponding probability of that storyline occurring.

Stochastic Weather Generator (SWG)

The SWG creates time series of synthetic weather for a given location based on the statistics of previously recorded measurements at that location. SWGs are commonly used to better characterise the weather in areas with limited historical observations, study the impacts of climate change on a specific area, and provide useful information for agricultural planning. Currently, the SWG designed for this project generates time series of daily precipitation, T_{max}, T_{min}, surface downwelling shortwave radiation (rsds) and potential evapotranspiration (PET), hereafter referred to as the climate variables, for any location for which data can be extracted from the VCSN data set. The SWG is designed to ensure that in any simulation of these climate variables, the time series are internally physically consistent with one another. The SWG can produce simulations for a single site or for several sites simultaneously.

This SWG is based on a standard Richardson style weather generator (Wilby 1999) and closely follows the design described in Parlange (2000). This kind of weather generator works in three key steps for each simulated day: (i) determine if it is raining; (ii) if it is raining, generate the amount of rainfall; (iii) calculate the other required simulation variables. By iteratively repeating this process, a coherent set of climate variable time series can be generated. In this section, the data preparation process and implementation of the stochastic weather generator are explained.

SWG training data

To simulate the climate variable time series listed above for a prescribed storyline, the SWG is provided with the detrended VCSN data series for that site for each day. These VCSN data are used to establish a statistical description of the time series as a basis for generating realistic, but synthetic, time series of the climate variables representative of the prescribed storyline.

Initialising the SWG

As the state of the model depends on the weater state on the previous day, to initiate the model, an 'initialisation' data set needs to be provided. To ensure that the initialisation is coherent between the different variables, a random day of data from the given month and weather state is used. These initialisation data are not included in the simulation output and are used only for initialisation purposes. While more sophisticated methods could be used to create these initialisation data, e.g., coherently extracting values from the distributions or using a real day and letting the generator spin up iteratively, the advantages of these more complex methods were found to be too small to warrant implementing. After the initial conditions have been determined, the main loop of the SWG is entered and all calculations are repeated iteratively, with each iteration corresponding to a single day.

Step 1 - Determine if current day is wet or dry

The SWG first determines whether it is raining, i.e., if it is a wet day or a dry day. To do this it first considers the weather state on the previous day and the length of the current precipitation streak, i.e., the number of previous consecutive wet or dry days, as this affects the probability of the current day being wet or dry. The SWG finds all historic streaks of the same length as the current streak and uses them to calculate the historic transition probability of the current streak being followed by today's state, be it wet or dry. If the transition probability is zero, i.e. no historic streaks of the specified length exist, the mean transition probability of all historic streaks is used. This mean transition probability is used as other approaches, such as using the probability from a shorter streak, can result in endless streaks. A random number between 0 and 1 is generated and, if higher than (1.0-PR_{dry}), where PR_{dry} is the probability of transitioning to a dry day, then the current day is set to being dry and otherwise is set to being wet.

Step 2 - If current day is wet, determine the precipitation

If it is determined that the current day is wet, the amount of rainfall on that day needs to be determined. This requires the cumulative distribution of precipitation values for the given month and weather state. For the purposes of generating these normalised cumulative distribution functions (CDFs) the VCSN data are segregated into just three precipitation states based on the historical classifications (please refer to the *"Storyline and state definition"* section of this document*)*:

1. **Wet** including the H/W, C/W and AT/W weather states.
2. **Dry** including the H/D, C/D, AT/D weather states.
3. **Average** including H/AP, C/AP, AT/AP weather states.

This coarser disaggregation into just three CDFs ensures that rarer states are still well represented. Once the CDFs have been generated, a random number between 0 and 1.0 is selected and, using linear interpolation, is used to derive a precipitation value. A limitation of this approach is that the SWG will not produce precipitation values greater than the largest value in the observational record. An approach based on fitting exponential functions was also developed and, while this more realistically captured extreme events, it produced worse results in general.

Step 3 - Calculate the other climate variables

With the precipitation time series simulated, the model generates compatible time series of the remaining climate variables. This process is non-trivial due to the interdependence of each of the variables. For example, a day with high T_{max} it is more likely to have a high T_{min}. Furthermore, there is strong temporal auto-correlation, e.g., a low T_{min} yesterday makes a low T_{min} today more likely. The Z-score for a variable *x* from a data set Iy is defined as:

$$Z = \frac{x - \bar{x}}{\sigma_x}$$

The SWG manages the intervariable and temporal dependencies by calculating such Z-scores for each variable for each day in a manner dependent on the Z-scores of all variables from the previous day and random perturbations to the previous day's variables. Using the Z-score instead of the variable simplifies the simulation by removing the need for more sophisticated weighting. The process of generating the climate variables in this manner is described in Wilby (1999) where the following equation is used:

$$Z_t = AZ_{t-1} + B\varepsilon_t$$

where Z_t is the vector of the Z-score values for each of the simulated variable values (T_{max}, T_{min}, rsds and PET) at time t. ε_t is a vector used to introduce the random forcing to the model and consists of a randomly generated value for each of the climate variables. The values for

ε_t are calculated using a multivariate normal distribution of a dimension equal to the number of variables in Z_t and with a mean of zero. It is easiest to think of the A and B matrices as a set of weights that describe how much each Z_{t-1} and ε_t value should impact Z_t. A detailed description of the calculation process is included in the appendix. As the A and B matrices have no time dependence, they are calculated before the SWG is run and are then reused on every iteration. As there is a dependence on Z_{t-1}, the generated Z_t must be stored and reused in the next iteration.

Once the methodology described above has been used to stochastically simulate the Z-scores, they then need to be converted back to physically relevant variables. While this could be done by fitting a theoretical distribution to the underlying data, this would require close inspection of every fit to ensure a robust result which is not possible given that the SWG is designed to work on any grid cell in the VCSN data set and with individual storylines. Instead, for each climate variable, the empirical CDF is used. The use of three CDFs, as detailed above, makes it possible to account for the effect precipitation has on the other variables. Each Z-score can be directly converted into a percentile and indexed to a value in the CDF.

As the relationships used by the SWG are purely statistical in nature, there is a small chance that a simulated T_{max} value will be smaller than the T_{min} value from the same day. The SWG is designed so that if this happens, the values will quickly correct and T_{max} will become greater then T_{min}. However, to avoid simulating this unphysical situation, if on any day T_{max} is simulated to be smaller than T_{min}, their values are swapped. The Z-scores are unchanged and the simulation continues as if the values were not swapped. While other approaches were tried for remedying this problem, this was selected as it produced the highest quality simulation.

Coherence across multiple sites

While the method described above generates a physically consistent set of climate variable time series consistent with the externally prescribed storyline, there may also be a need to generate such sets of time series for several sites. It is not sufficient to simply run the model again at the new location as, Given the random elements of the SWG, the spatial coherence of the time series would be lost, i.e. a site with an anomalously high T_{max} would be expected to have nearby sites also experiencing anomalously high T_{max} values. To capture this spatial coherence, once the climate variable time series for the primary site have been generated, time series for a secondary site are simulated as follows:

- As in the case for a single site, the SWG first determines if it is raining at the secondary site. To do this, the SWG uses the probability of it raining at the secondary site given the rain/no rain state at the primary site. This results in there still being an indirect dependence on the current wet or dry streak through the primary site.
- Next, if it is determined to be raining, the precipitation amount is again determined by selecting a random number between 0 and 1.0 and linearly interpolating within the normalised CDF. However, instead of using the CDF of precipitation, the CDF of the *difference in precipitation* between the two sites is generated and used, where the two locations are within the same climate state and month, and given that it rained at both locations. This difference value is then added to the primary site's simulated

value to give the precipitation at the secondary site. If this results in a negative precipitation value at the secondary site, the value is multiplied by -1. The same approach is used for the other climate variables.

This approach ensures that the spatial coherence between the climate variable time series at the primary and secondary site is preserved. If, however, the intention is to use the SWG to generate climate variable simulations for 3 or more sites, the coherence between the second and third site cannot be assured. For example, given a primary site *A*, and two secondary sites *B* and *C* whose time series were both 'locked' to those at site A by way of the method described above, while coherence between the time series at sites *A* and site *B* is captured, and the coherence between site *A* and site *C* is captured, there is no guarantee that the coherence between sites *B* and *C* is simulated accurately.

References

Black, M.T.; Karoly, D.J.; Rosier, S.M.; Dean, S.M.; King, A.D.; Massey, N.R.; Sparrow, S.N.; Bowery, A.; Wallom, D.; Jones, R.G.; Otto, F.E.L. and Allen, M.R., The weather@home regional climate modelling project for Australia and New Zealand, Geoscientific Model Development, doi:10.5194/gmd-9-3161-2016, 2016.

Parlange, M. B., & Katz, R. W. (2000). An Extended Version of the Richardson Model for Simulating Daily Weather Variables. *Journal of Applied Meteorology*, 39, 610–622. https://doi.org/10.1175/1520-0450-39.5.610

Tait, A.B. and Turner, R., Generating multiyear gridded daily rainfall over New Zealand, *Journal of Applied Meteorology*, 44, 1315–1323, 2005.

Tait, A.B.. Future projections of growing degree days and frost in New Zealand and some implications for grape growing, Weather and Climate, 28, 17–36, 2008.

Wilby, R. L. (1999). The weather generation game: A review of stochastic weather models. *Progress in Physical Geography*, *23*(3), 329–357. https://doi.org/10.1177/030913339902300302

Appendix

A and B Matrix Calculation

The A and B matrices are calculated following the process described in Parlange (2000) and summarised below:

- First the matrices m_0 and m_1 (the lag 0 and lag 1 cross covariance matrices) need to be calculated. The following equations are used, where $corr(x, y)$ is the pearson correlation coefficient between variables x and y:
    - m_0 consists of values $p_{k,\,l} = corr(\ Z_t(k),\ Z_t(l)\ )$
    - m_1 consists of values $p_{k,\,l} = corr(\ Z_t(k),\ Z_{t-1}(l)\ )$
- These values can be used to estimate the autoregression matrix, A, and the variance-covariance matrix S. This is done by solving the equations $Am_0 = m_1$ and $S = m_0 - Am_1{}^T$
- Once A has been determined, B can be calculated to run the simulations. B is related to S through the equation $S = BB^T$. Given the properties of the S matrix, this equation can be solved to give B through the use of cholesky decomposition.