A New Quantile Projection Method for Producing Representative Future Daily Climate based on Mixed Effect State-Space Model and Observations

Huidong Jin[1], Phil Kokic[1], Garry Hopwood[2], J.H. Ricketts[3] and Steven Crimp[2]

[1]*CSIRO Digital Productivity Flagship, GPO Box 664, Canberra ACT 2601, Australia.*
Email: Warren.Jin@csiro.au..
[2]*CSIRO Agriculture Flagship, GPO Box 1700, Canberra ACT 2601, Australia*
[3] *Department of Science, Information Technology and Innovation (DSITI), Dutton Park, Qld, Australia*

Abstract: Agriculture is particularly sensitive to climate variability and change. In Australia, agricultural production spans the tropics, sub-tropics, Mediterranean and temperate climatic zones, and operates under conditions of high climatic variability compared with other countries (Stokes & Howden, 2010). In response to operating under these conditions good risk managers have been actively using climate variability information to support operational management decisions. Increasingly these producers are now considering the longer term implications of anthropogenic climate change in their strategic farm planning (Meinke & Stone, 2005). To meet these longer term requirements, the Queensland Government through its Queensland Climate Change Centre of Excellence (QCCCE, now the Department of Science, Information Technology and Innovation (DSITI)), started delivering climate change projection data in 2010 via the Consistent Climate Scenarios (CCS) data portal (Burgess *et al.*, 2012). Since mid-2012, CCS projections data have been more widely available via the CCS web-portal on DSITI's Long Paddock website, providing representative point-referenced *daily* data for periods centred on 2030 and 2050. Currently, registered users can access projected daily climate data free of charge as patched-point data (for any of the 4,759 SILO weather stations), or as individual points from a 0.05 degree grid across Australia. The future projections are based in part on observed trends as well as from Global Climate Models (GCM) and emission scenarios from the Intergovernmental Panel on Climate Change (IPCC), Fourth Assessment Report (AR4). CCS data are available in formats suitable for most biophysical models, enabling various stakeholders to undertake climate change adaptation studies for various primary industries across Australia, in particular within the grazing, cropping and horticultural sectors. To generate plausible and weather-like point-referenced future daily data from coarse-scale GCM outputs (typically monthly values on a 100-200km spatial grid), two downscaling approaches were developed, change factors and quantile matching. The quantile matching employs regression on monthly quantiles from daily observations to project future quantiles; and spatial pattern scaling to adjust quantile means based on GCM outputs. Quantile matching is used to simulate future daily climate based on historical observations, which incorporates quantile forecast bias correction, auto-correlation and cross-variable correlations.

The CCS system has recently been refined to incorporate new GCM outputs from the IPCC Fifth Assessment Report (AR5), as outcome from a DSITI and CSIRO collaboration. Improvements on the quantile projection method are described in this paper. We use a multivariate Linear Mixed Effect State Space (LMESS) model for all the possible historical data to replace simple linear regression on data from each single month (Kokic *et al.*, 2011). We model 12 months of data and three quantiles together to handle missing data issues in the historical observations. We select the same set of covariates from time and seasonality terms in order to maintain the consistent cross-variable correlations among the climate elements. We allow time-varying coefficients for some covariates. The model parameters are estimated via an Expectation-Maximisation algorithm. It is implemented in the statistical computing language R, and is called by the CCS from Python via the package rpy2. The model was tested and validated on dozens of representative SILO stations across the Australian continent. The hindcast results show that LMESS improves the hindcast accuracy and maintains the quantile spread, over the monthly regression used in the CCS system, linear regression (LR) and multivariate LR. Whilst the improvements are not statistically significant, the results are sizeable enough to support a move to this revised modelling approach.

Keywords: *Bootstrapping, climate change projections, consistent climate scenarios, LMESS, quantile matching*

1. INTRODUCTION

In response to operating in a highly variable climate and under the longer-term implications of climate change (Stokes & Howden, 2010), the Queensland Government through its Queensland Climate Change Centre of Excellence (QCCCE, now the Department of Science, Information Technology and Innovation (DSITI)), started delivering climate change projection data via the Consistent Climate Scenarios (CCS) project (Burgess et al., 2012). Since mid-2012, CCS projections data have been more widely available via the CCS web-portal on DSITI's Long Paddock website. It provides representative point-referenced *daily* data for periods centred on 2030 and 2050. CCS data include six climate elements including rainfall, evaporation, maximum and minimum temperature, and solar radiation. These daily data are available in formats suitable for most biophysical models, enabling various stakeholders to undertake climate change adaptation studies for various primary industries across Australia; in particular within the grazing, cropping and horticultural sectors.

To generate plausible and weather-like point-referenced future daily data from coarse-scale GCM outputs (typically monthly values on a 100-200km spatial grid), two downscaling approaches were used in the CCS, i.e., Change Factors (CF) and Quantile Matching (QM) (Ricketts *et al.*, 2013). CF first uses a linear model to compute the linear trend between the global annual average surface temperature and the mean monthly values of a climate variable for all the GCM grid points, and these trends are then interpolated into a finer resolution to form a pattern of change. These spatial patterns are then used to generate change factors by multiplying the pattern of change by the mean global temperature projected at a future time-point (e.g. 2030). QM considers four targets: consistency with the historical observations; consistency with the broader scale climate data used for the projections; the degree of transparency and repeatability; and the plausibility of results produced (Kokic *et al.*, 2013). It employs a linear regression model on monthly quantiles (i.e., 10th, 50th and 90th percentiles in CCS) from daily observations to project future quantiles for each month (we call it monthlyLM). Spatial pattern scaling is used to adjust quantile means based on GCM outputs and at the same time maintain the spread of quantiles, such as differences between the 90th (or 10th) percentiles and 50th percentiles. These quantiles are used to form simple cumulative distribution functions (CDF) for each month. For a historical daily observation, we calculate the quantile under the historical monthly CDF, and the value corresponding to this quantile under the future CDF is viewed as a representative future daily climate. QM in the CCS incorporates quantile forecast bias correction, auto-correlation and cross-variable correlations among the climate elements as detailed in (Kokic *et al.*, 2013). monthlyLM forecasts quantiles for five climate elements: rainfall, radiation, humidity, minimum and maximum temperatures. To satisfy modelling assumption, climate elements may need to be pre-transformed, e.g. such as cubic root for rainfall.

There are several issues with the quantile forecast method monthlyLM in QM (as well as CF). For some stations and months, there are many missing quantiles, e.g rainfall quantiles for Mays for Station 2038. The lack of monthly level data make it impossible to estimate the linear trend against time in QM or global annual average surface temperature in CF. Secondly, monthlyLM does not consider the possible dependence among quantiles, which often leads to quantile crossover (e.g., the 90th percentiles become smaller than 10th percentiles after 2017 in the left panel of Figure 1). Thirdly, to handle the lower reliability of monthlyLM for long term forecast, there are two variations for QM. The QM 2050 projections requires GCM daily data which are not normally available. This introduced unnecessary complexity.

To handle these issues, a new quantile forecasting method was implemented. At the same time, most components in the current CCS system were reused. The new approach involves using a Linear Mixed Effect State Space (LMESS) model. It models 12 months of data together to handle missing data issues. To maintain consistent distributional characteristics from historical observations, the approach has been implemented to model three quantiles and their covariances simultaneously. It uses one identical set of covariates to maintain consistent cross-variable correlations among climate elements. It also supports dynamic coefficients for some covariates to enhance forecast accuracy.

We will discuss our modelling procedure and four different models in Section 2. After implementation description in Section 3, various comparison results are presented in Section 4. In Section 5, we provide some concluding comments and possible future development ideas.
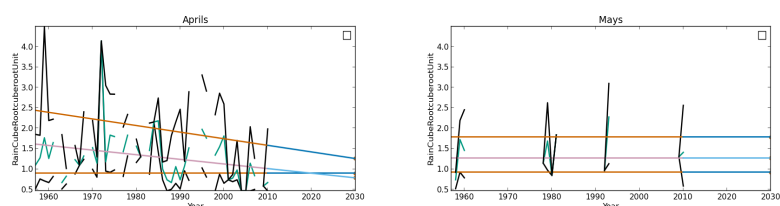

Figure 1. Issues of monthly linear model for quantile crossover (left panel) and lack of data (right) for rainfall of Station 2038

2. QUANTILE MODELLING AND FORECAST

To handle the quantile forecast issues in the current CCS system, we will fit and forecast the three quantiles for one climate variable simultaneously. To alleviate the data shortage issue for some months, we will model 12 month data, instead of single months. As the GCM output will be used in the later stage, we use covariates only

Figure 2. Quantile forecast from LMESS to handle issues like quantile crossovers and lack of data for rainfall of Station 02038

related with time, such as year, month, and seasonality terms. Furthermore, in order to maintain the spatial coherence among stations to some degree (Kokic *et al.*, 2013), we should use the same set of covariates for all the stations. To have one model for both forecast time points 2030 and 2050, we will also allow a dynamic coefficient for some covariates. All these goals are achieved using LMESS.
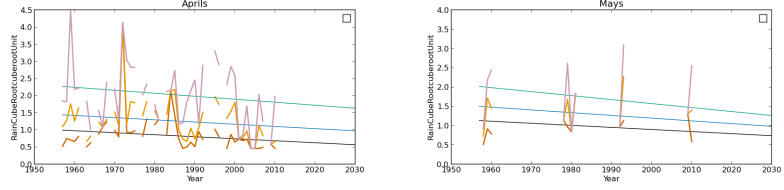
2.1. Modelling and forecast

As the linear regression and multivariate linear regression are two special cases of LMESS, we briefly describe the model proposed by Kokic *et al.* (2011) first.

Assume we have daily measure of the climate element, say daily temperature, from 1950 to a time point T, say, 2010. Following the QM in the CCS, we first estimate r (=3) quantiles as the summary statistics for each month. These monthly quantiles, denoted by $\tilde{q}_{1t}, \tilde{q}_{2t}, \cdots, \tilde{q}_{rt}$ ($\in$ R), are necessary for modelling the distribution of the daily climate elements. Also assume there are covariates $\mathbf{u}_1$, $\mathbf{u}_2$, and so on. The quantile forecast procedure aims to forecast these statistics to some future time point $T_F > T$ using the covariate information up to time T_F. Instead of using output from GCMs as covariates in (Kokic *et al.*, 2013), we only use time-related terms as our covariates. Following the notation in Kokic *et al.* (2011), after pre-defined transformation of $\tilde{q}_{1t}, \tilde{q}_{2t}, \cdots, \tilde{q}_{rt}$, we denote these quantiles by $\mathbf{y}_t = (y_{1t}, \ldots, y_{rt})^T$. Pre-transformations can be used to improve the fit of the model to the data. The LMESS model is specified in terms of an observation equation and system equation. The observation equation is of the form:

$$\mathbf{y}_t = C_t \alpha_t + D_t \beta + v_t \tag{1}$$

where $\alpha_t = (\alpha_{1t}, \cdots, \alpha_{kt})^T$ are unobserved state variables (random effects), C_t is an r by k matrix of possibly unknown values, $\beta = (\beta_1, \cdots, \beta_m)^T$ are unknown parameters representing the fixed effects in the model, D_t is an r by m matrix of known values, and v_t is a vector of observation errors, assumed to be multivariate normally distributed with mean $\mathbf{0}$ and variance $R \in R^{r \times r}$. In addition, the α_t variables are assumed to be random and evolve over time according to the following system equation:

$$\alpha_t = A\alpha_{t-1} + w_t \tag{2}$$

where A is a $R^{k \times k}$ matrix of possibly unknown parameters and $\mathbf{w}_t$ is a multivariate normally distributed error term with mean 0 and variance $Q \in R^{k \times k}$. In addition, $\mathbf{v}_t$ and $\mathbf{w}_t$ are assumed to be serially uncorrelated and uncorrelated with each other. Furthermore, for simplicity, R and Q are assumed to be positive definite.

The multivariate linear regression (MLR) is a special case of the LMESS model with no state variables and without the C_t matrix. One advantage of the LMESS model over MLR is that it allows for the possibility of serial correlation of the *y*-values, which is a common feature of climate data. Linear Regression (LR) over a single quantile is a further special case of LMESS and MLR with a diagonal covariance matrix R. In our application, both C_t and D_t are design matrices based on the covariate information $\mathbf{u}_1$, $\mathbf{u}_2$, and so on.

Based on the assumption that the conditional (prior) distribution of α_{t-1} given $\mathbf{y}_{t-1}$ is normal, $(\alpha_{t-1} \mid \mathbf{y}_{t-1}) \sim N(\tilde{\alpha}_{t-1}, \tilde{\Sigma}_{t-1})$, using the Kalman filter and the EM algorithm, we can obtain maximum likelihood estimates of the unknown parameters such as $\tilde{\alpha}_t, \tilde{\Sigma}_t$ (for t=1,…,T), β, A, R and Q. When forecasting beyond T, the state variable α_t has a distribution conditional on all the observations $\mathbf{y}_{1,\cdots,T}$ up to time T, $(\alpha_t \mid \mathbf{y}_{1,\cdots,T}) \sim N(\tilde{\alpha}_t, \tilde{\Sigma}_t) = N(A\tilde{\alpha}_{t-1}, A\tilde{\Sigma}_{t-1}A^T + Q)$ for $T < t \le T_F$. Similarly, the forecast $\mathbf{y}_t$ follows a

conditional distribution $(\mathbf{y}_t \mid \mathbf{y}_{1,\cdots,T}) \sim N(C_t\tilde{\alpha}_t + D_t\beta, C_t^T\tilde{\Sigma}_tC_t + R)$. Removing the fixed effect term $D_t\beta$ in this forecast equation, it is identical with the one in the classic state-space model (Kokic *et al.*, 2011).

2.2. Covariates and selection

To comply with QM in the CCS system, we restrict ourselves to the use of year and month, seasonality terms and their interactions as our covariates. The covariates we used include timestamp (year+(month-1)/12), seasonality terms like sin(month/12*2π*c)+cos(month/12*2π*c), for $c = 1,2,\cdots,6$. The constant c indicates how many cycles within one year. For simplicity, we denote sin(month/12*2π*c) and cos(month/12*2π*c) as sin*c* and cos*c* respectively. The interaction terms include year×sin1 and year×cos1.

We may use the variable selection mechanism built in LMESS or linear regression to select covariates for each climate variable and each station on the fly. However, possibly different subsets of covariates may introduce unnecessary inconsistency during the forecast. In order to maintain the cross-variable consistency within a station and cross-station consistency, we use the same subset of covariates for all stations. As a quick solution, we use linear regression on all the climate variables of interest against a set of covariates. We then count the number of a covariate having a significant coefficient for a given significance level. The covariates that have a higher significant frequency are selected as covariates in the final quantile forecast.

3. IMPLEMENTATION, TEST AND EXPERIMENTAL SETTING

The quantile forecast models were implemented in the programming language R. We also used the EM algorithm which can handle missing data in multivariate state-space models (Kokic & Jin, 2015). The EM algorithm would output parameter estimates when a maximum iteration number, like 100, is reached or the relative likelihood change becomes very small, e.g. less than 10^{-5} between two consecutive iterations. To enhance the robustness of the quantile forecast part, if the EM algorithm fails, e.g., for inappropriate input data, the R code catches the exception and switches to a linear regression model to forecast quantiles. The CCS system is primarily implemented in another programming language Python. We used the python package rpy2 to communicate with R code.

We selected 18 geographically dispersed stations which represent different climatic zones in Australia to assess the performance of the approach across different climatic conditions. We used hindcast to compare different models in order to provide evidence on the most suitable model. For the 18 stations, our code has been tested on both Windows and Linux platforms, and CSIRO and DSITI servers

4. MODELLING RESULTS AND COMPARISON

4.1. Subset of covariates selected

We first report the subset of time-related covariates we chose. As discussed above, we use linear regression of a quantile against all the possible covariates. For a given significance level, we calculate the frequency each covariate is significant in these 810 LR models.

First, sin5 and cos5, and sin6 and cos6 are rarely statistically significant, and they are dropped first. From the left panel of Figure 3, we can see sin4 and cos4 are not often statistically significant (as their frequencies are less than 20% or 10% for a significance level 0.05 and 0.01, respectively). The interaction terms, as well as sin1 and cos1 are slightly more frequent. After removing the two interaction terms (right panel, Figure 3), we can see sin1 and cos1 are really important. Finally, we chose timestamp, sin1, cos1, sin2, cos2, sin3 and cos3 as covariates.
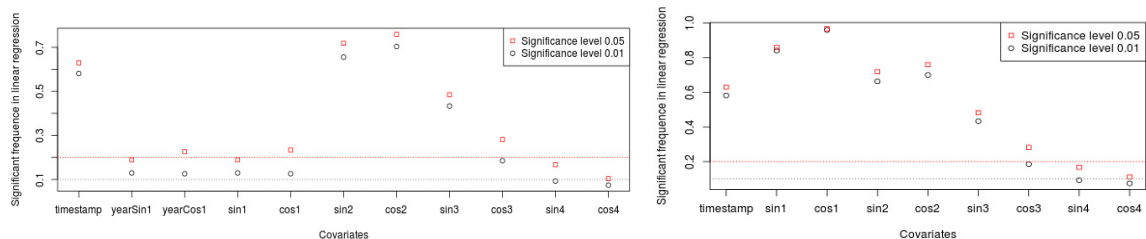

Figure 3. Significant frequency of covariates in various linear regression for different significance level (the left panel has two interaction terms between time and seasonality)

4.2. Cross validation and comparison

We conducted hindcasts to evaluate the four models: LMESS, MLR, LR and MonthlyLM. We used the 10th, 50th and 90th percentiles calculated from observations before 2009 to fit the models and evaluated the models' forecast against quantiles between 2009 and 2010. The validation metrics focus on the predictive accuracy and the accuracy on maintaining the spread between quantiles (90th percentile - 50th percentile and 50th percentile – 10th percentile). For each station, we have 72 validation values for each climate element. For predictive accuracy, the predictive mean square error (predMSE) measure of the mean difference between the modelled quantiles and the observed quantiles. The predictive standard deviation (predSD) indicates the variation of these predictive errors. For the differences among quantiles, we used similar metrics to estimate the mean square errors and standard deviations, diffPredMSE and diffPredSD. For LMESS, we treated 'timestamp' with a dynamic coefficient. We also tried to treat 'intercept' as a random effect, but that did not generate much more accurate results in terms of predMSE and diffPredMSE. So, only 'timestamp' was modelled as a random effect in the final LMESS setting.

Figure 4 illustrates the fitted and forecasted 50th percentiles from monthlyLM (top panel) and the LMESS (bottom panel) for median rainfall (which is under the cube root transformation) for Station 2038. Both models fit the quantiles reasonably well, while LMESS can follow some temporal changes much better during the fitting period from 1955 to 2008. The monthlyLR approach results in two rapid changes within each year, which is not necessarily true. In addition, these rapid changes led to the issue of quantile crossovers in Figure **1**. LMESS gives a relatively smooth forecast, without abrupt intra-annual changes. The predictive MSE is smaller for the LMESS approach, i.e., 0.432 versus 0.467 and delivers a 7% improvement in predictive accuracy on monthlyLM based on the hindcast results. For this station, predMSE=0.529, predSD=0.497, diffPredMSE=0.458, diffPredSD=0.434 for monthlyLM. predMSE=0.474, predSD=0.472, diffPredMSE=0.438, diffPredSD=0.441 for LMESS. LMESS outperforms monthlyLR in terms of both predMSE and diffPredMSE.
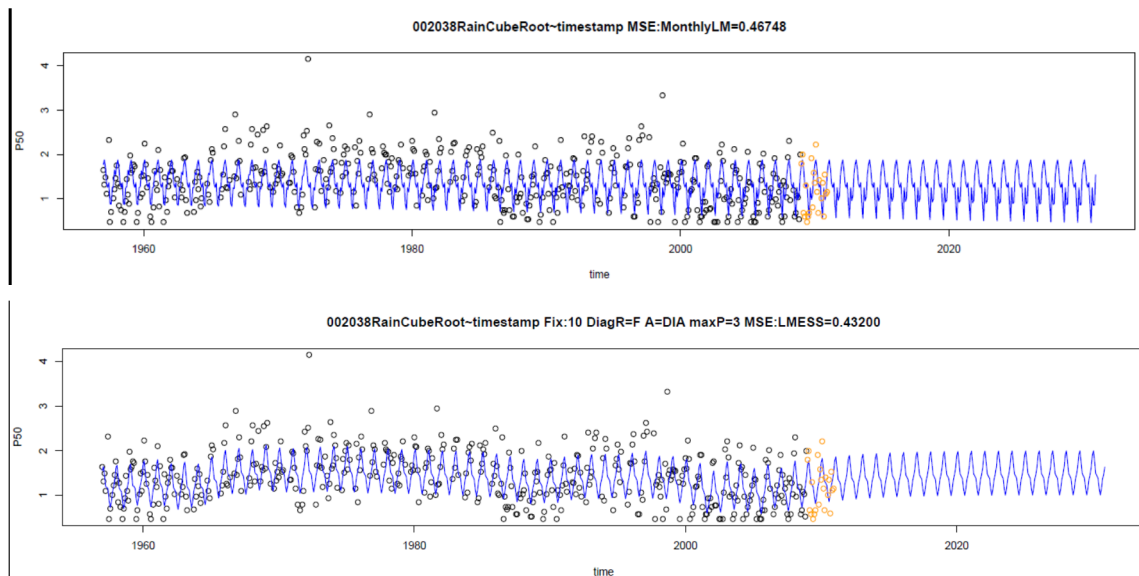

Figure 4. Fitted and forecast 50th percentiles from MonthlyLM (top) and LMESS (bottom) (dots are percentiles from observations, black dots for training and orange dots during 2009-2010 for validation)

Figure 5 to Figure 7 show boxplots of one accuracy measure of four different models over 18 stations. For each boxplot, the bottom and top of the box are the first and third quartiles, and the band inside the box is the median value. The standard Tukey method (i.e., $1.5 \times$ IQR) is used for the whiskers. The white dot within a boxplot indicates the mean value.

In terms of mean value of predMSE in Figure 5, monthlyLM often has slightly higher value. On average, it has a higher predMSE on humidity, radiation, and rainfall than the other three models. It has higher predMSE than maximum and minimum temperatures than LMESS. LMESS normally has lower mean predMSE, especially compared with the other three on minimum temperature and radiation. MLR has similar average performance as LMESS on humidity, maximum temperature, and rainfall, while slightly worse on minimum temperature and radiation. LR has a quite similar performance as MLR, except it is worse for minimum temperature.

In terms of predictive variability in Figure 6, on average, LMESS has slightly lower value than the other three models over five different climate elements. MonthlyLM has a slightly larger value among the four models. MLR and LR sit in between. They perform as well as LMESS on humidity and rainfall, while slightly worse on the other three climate elements.
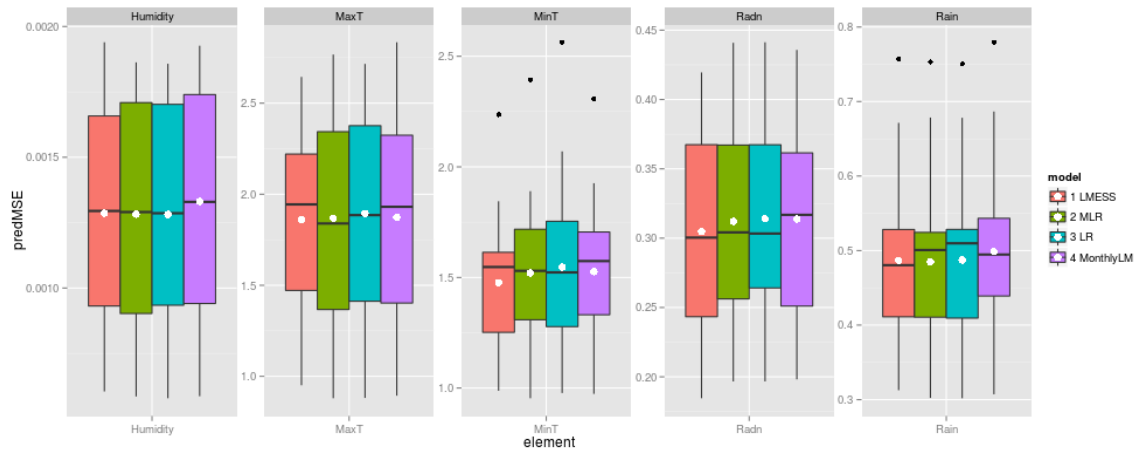

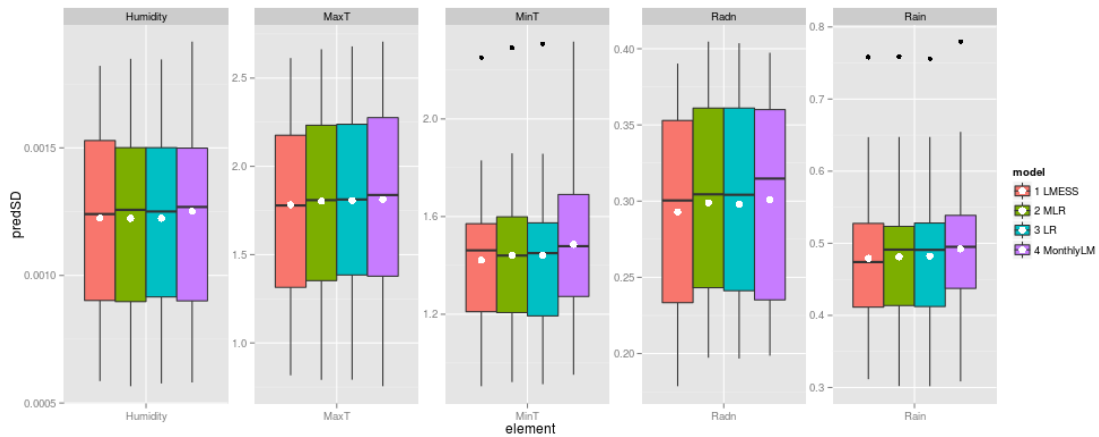**Figure 5.** Predictive MSE on 3 quantiles of four models over 18 stations for five climate elements

Figure 6. Predictive standard deviation of errors on quantiles of four models over 18 stations

Similar comparison results could be found in their ability to maintain the difference between quantiles. On average diffPredMSE (Figure 7) and diffPredSD, LMESS has the best performance on radiation, maximum and minimum temperatures. MonthlyLM has the worse forecast performance on all the five climate elements. LR and MLR lie in between, though they are slight better than LMESS on modelling rainfall.

Based on hindcast results, the forecast accuracy improvement is small for LMESS against the other three models, for example only 0.05°C improvement on minimum temperature. However, considering the cross-validation is against just 1 year or 2 years after the training data, we expect the performance gap to become larger when forecasts are made for 20 or 30 years into the future.

5. CONCLUSIONS AND DISCUSSIONS

For the consistent climate scenario project, we have proposed a new quantile forecast method based on LMESS in order to improve performance. Our hindcast results against observations in 2009-2010 have shown that the monthlyLM performs reasonably well though sometimes it has some issues on handling data shortage and quantile crossovers. In terms of predictive accuracy and maintaining the quantile spread, LMESS normally generates more accurate quantile hindcast results than monthlyLM in the old CCS system, as well as the two special cases of LMESS, linear regression (LR) and multivariate LR (MLR). The improvement is not statistically significant, but we hope the improvement might become bigger for forecasts with a longer time horizon. We suggest the use of LMESS in the refined CCS system. Considering the

computation and modelling complexity of LMESS, linear regression (using all the 12 months data) is a good alternative.

In the CCS, we ignore the dry days when modelling rainfall which leads to some data shortcomings for some dry months. Modelling the wet day proportion in LMESS could improve the forecast accuracy, especially where there is an increasing or decreasing trend of wet days in the future. Another possible extension is to consider spatial dependency among stations using spatio-temporal statistical models like spTDyn (Bakar *et al.*, 2015a, b).
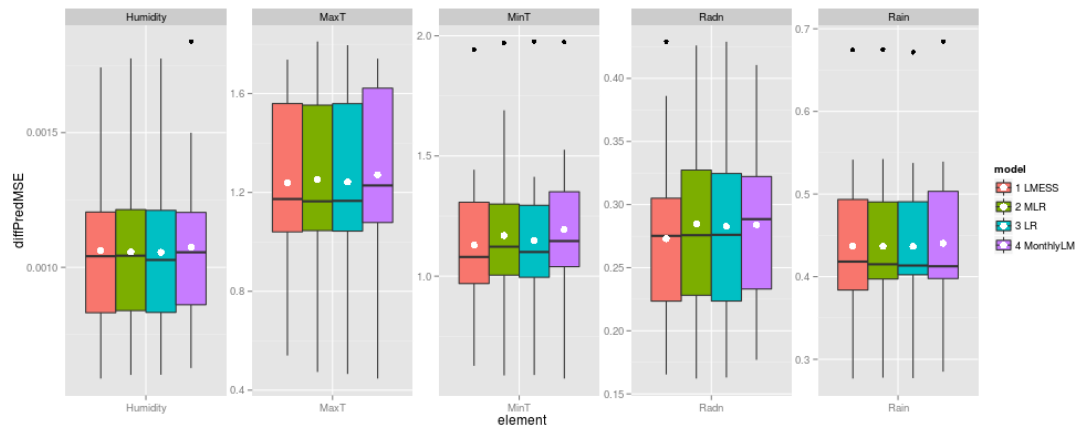

Figure 7. Predictive standard deviation of errors on quantile differences of four models over 18 stations

REFERENCES

Bakar, K.S., Kokic, P. & Jin, H. (2015a) A spatiodynamic model for assessing frost risk in south‐eastern Australia. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, Accessible online

Bakar, K.S., Kokic, P. & Jin, H. (2015b) Hierarchical spatially varying coefficient and temporal dynamic process models using spTDyn. *Journal of Statistical Computation and Simulation*, 1-24, In Press.

Burgess, S., Ricketts, J., Panjkov, A., Carter, J.O. & Day, K. (2012) Consistent Climate Scenarios Project Draft User Guide: 'Change factor' and 'Quantile-matching' based climate projections data In. The Department of Science, Information Technology, Innovation and the Arts (DSITIA), Queensland Government, https://www.longpaddock.qld.gov.au/climateprojections/guide.html.

Kokic, P. & Jin, H.D. (2015) Parameter estimation using the EM algorithm for multivariate state-space models with missing data. *TBA, Under review*,

Kokic, P., Crimp, S. & Howden, M. (2011) Forecasting climate variables using a mixed-effect state-space model. *Environmetrics*, **22**, 409-419.

Kokic, P., Jin, H.D. & Crimp, S. (2013) Improved point scale climate projections using a block bootstrap simulation and quantile matching method. *Climate Dynamics*, **41**, 853-866.

Meinke, H. & Stone, R. (2005) Seasonal and inter-annual climate forecasting: The new tool for increasing preparedness to climate variability and change in agricultural planning and operations. *Climatic Change*, **70**, 221-253.

Ricketts, J., Kokic, P. & Carter, J. (2013) Consistent Climate Scenarios: projecting representative future daily climate from global climate models based on historical climate data. In: *MODSIM2013, 20th International Congress on Modelling and Simulation* (ed. J.a.A. Piantadosi, R.S. And Boland J.), pp. 2785-2791. Modelling and Simulation Society of Australia and New Zealand

Stokes, C. & Howden, M. (2010) *Adapting agriculture to climate change: preparing Australian agriculture, forestry and fisheries for the future*. CSIRO publishing.