Spatial modelling of linear regression coefficients for gauge measurements against satellite estimates

Benjamin Hines, Yuriy Kuleshov and Guoqi Qian

Abstract Satellite imagery provides estimates for the amount of precipitation that has occurred in a region, these estimates are then used in models for predicting future precipitation trends. As these satellite images only provide an estimate for the amount of precipitation that has occurred, it is important that they be accurate estimates. If we assume that a rain gauge correctly measures the amount of precipitation that has occurred in some location over a specified time interval, then we can compare the satellite precipitation estimate to the gauge measurement for the same time interval. By expressing the relationship between the gauge measurement and the satellite precipitation estimate for the same time interval as a linear equation we can then spatially map the coefficients of this linear relationship to inspect the spatial trends of the regression coefficients. We then model the coefficients of the linear equations of each location by a spatial linear model and then use this model to predict the coefficients in location where there are no rain gauges available.

1 Introduction

The ability to measure precipitation accurately is very important for many reasons. Knowing how much precipitation has occurred in a given region will help us improve our knowledge about the seasonal patterns and climate trends in said region and the regions around it. Rain gauges are used to measure the amount of precipitation that has occurred in a given time period, where precipitation is captured and then measured at equally spaced intervals. Figure 1 shows the locations of all the

Benjamin Hines and Guoqi Qian
School of Mathematics and Statistics, The University of Melbourne, Parkville VIC 3010, Australia, e-mail: benjamin.hines@unimelb.edu.au, e-mail: g.qian@ms.unimelb.edu.au.

Yuriy Kuleshov
Australian Bureau of Meterology, Docklands Melbourne VIC 3008 Australia; School of Science, RMIT University, e-mail: yuriy.kuleshov@bom.gov.au.

Bureau of Meteorology's monthly rain gauges around Australia that are operational as of August 2018 in which there are 865 of. Clearly, the gauges are not uniformly placed around Australia, but are distributed based on the population concentration. If we wanted to know how much precipitation has occurred in a region where there

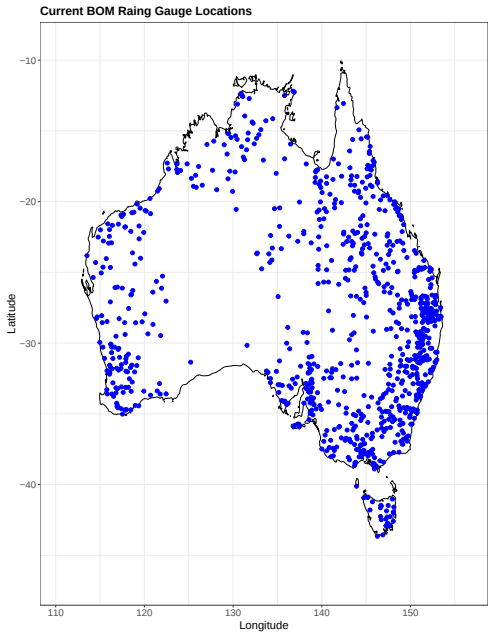

Fig. 1: Bureau of Meteorology's Australian monthly rain gauge locations as of August 2018.

are no rain gauges, we would have to rely on satellite precipitation estimates. Satellite precipitation estimation is done by taking an image of a cloud from above and then estimating the amount of precipitation that will come from that cloud based on its physical characteristics [9]. As these satellite images only provide an estimate for how much precipitation has occurred in a region, it is important to know if these estimates are close to what a rain gauge would measure. If the satellite images do indeed provide us with good estimates for the amount of precipitation, then there is no issue with using these estimates as measurements for regions in which there are no rain gauges located. However, if the satellite images do not provide us with a good estimate for how much precipitation a rain gauge has measured for some time interval for a given location, then we can look at the difference between the gauge measurement and the satellite estimate and how there may be some relationship between the two, which may depend on the location of interest.

2 Methodology

Let $\mathbf{x}_i \in \mathbb{R}^2$ for $i = 1, \ldots, m$ be the two dimensional coordinates description of the i^{th} location where x_{i1} and x_{i2} are the longitude and latitude for location i respectively. We then define $Y_{ij}^{[g]}$ and $Y_{ij}^{[s]}$ to be the gauge measurement and satellite estimate for the i^{th} location respectively for the j^{th} time period where $j = 1, \ldots, n_i$. We can then consider $Y_{ij}^{[g]}$ and $Y_{ij}^{[s]}$ to have a linear relationship, i.e.

$$Y_{i,j}^{[g]} = \beta_0^{[i,j]} + \beta_1^{[i,j]} Y_{i,j}^{[s]}.$$

Now if we consider the relationship between the gauge measurements and satellite estimates to be temporally stationary (coefficients are the same regardless of time) [5], then we can express a linear equation for location i as

$$\mathbf{Y}_i^{[g]} = \beta_0^{[i]} \mathbf{1}_{n_i} + \beta_1^{[i]} \mathbf{Y}_i^{[s]} + \boldsymbol{\varepsilon}_i \tag{1}$$

where $\mathbf{Y}_i^{[g]} = \left(Y_{i,1}^{[g]}, \ldots, Y_{i,n_i}^{[g]}\right)^T$, $\mathbf{Y}_i^{[s]} = \left(Y_{i,1}^{[s]}, \ldots, Y_{i,n_i}^{[s]}\right)^T$, $\boldsymbol{\varepsilon}_i \sim \mathcal{N}\left(\mathbf{0}, \sigma_i^2 I_{n_i}\right)$ is the noise term and $\mathbf{1}_{n_i}$ is an $n_i \times 1$ vector with every entry being 1. We would expect to estimate $\beta_0^{[i]} = 0$ and $\beta_1^{[i]} = 1$ for $i = 1, \ldots, m$ as we expect the satellite precipitation estimate to be equal to the gauge measurement. The conventional method for estimating the coefficients $\beta_0^{[i]}$ and $\beta_1^{[i]}$ in this situation is by minimising the ordinary least squares equation

$$\hat{\beta}^{[i]} = \underset{\beta_0^{[i]}, \beta_1^{[i]}}{\arg\min} ||\mathbf{Y}_i^{[g]} - \beta_0^{[i]} \mathbf{1}_{n_i} - \beta_1^{[i]} \mathbf{Y}_i^{[s]}||_2^2 \tag{2}$$

where $\hat{\beta}^{[i]} = \left(\hat{\beta}_0^{[i]}, \hat{\beta}_1^{[i]}\right)^T$ and $||\mathbf{t}||_p = (\sum_{i=1}^n |t_i|^p)^{1/p}$. The solution to equation (2) can be shown to be

$$\hat{\beta}^{[i]} = \left(V_i^T V_i\right)^{-1} V_i^T \mathbf{Y}_i^{[g]}$$

with $V_i = \left(\mathbf{1}_{n_i}, \mathbf{Y}_i^{[s]}\right)$ an $n_i \times 2$ matrix [3]. In doing this we can obtain estimates for the coefficients $\beta_0^{[i]}$ and $\beta_1^{[i]}$ for all locations which we define by

$$\beta_\ell = \left[\hat{\beta}_\ell^{[1]} \ \hat{\beta}_\ell^{[2]} \ \cdots \ \hat{\beta}_\ell^{[m]}\right]^T \tag{3}$$

an $m \times 1$ vector for $\ell = 0, 1$.

Once we have our estimates for the coefficients of each location as in equation (3), we can think of these spatially specified coefficients as a spatial process of a geostatistical random field [14], where we can construct a model to test if there is some sort of spatial dependency structure. Consider the following spatial linear model:

$$\beta_\ell = \alpha_\ell \mathbf{1}_m + \lambda_\ell W_{m,\ell} \beta_\ell + \varepsilon(\beta_\ell) \tag{4}$$

for $\ell = 0, 1$, where λ_ℓ is the autocorrelation parameter, α_ℓ is the intercept coefficient of the model, $W_{m,\ell}$ is a given $m \times m$ weight matrix representing the spatial distances of the observations and $\varepsilon(\beta_\ell) \sim \mathcal{N}\left(\mathbf{0}, \sigma^2(\beta_\ell) I_m\right)$ for $\ell = 0, 1$ [13]. Furthermore, $W_{m,\ell} = \left\{ w_{ij}^{[\ell]} \right\}$, where $w_{ij}^{[\ell]}$ is given by some function of a known distance metric $d(\mathbf{x}_i, \mathbf{x}_j)$ where locations $\mathbf{x}_i$ and $\mathbf{x}_j$ are the locations for the coefficients $\hat{\beta}_\ell^{[i]}$ and $\hat{\beta}_\ell^{[j]}$ respectively. Note that $w_{ii}^{[\ell]} = 0$ for all $i = 1, \ldots, m$ as the observation cannot depend on itself. The choice of which metric and function we use to define out weight matrix $W_{m,\ell}$ is vital to being able to model the spatial data well. If the weight matrix does not give a good representation of the true nature of spatial region of interest, then the estimates calculated for the parameters are likely to be biased and inconsistent [2, 6]. Clearly, good selection of the weight matrix is essential for providing unbiased estimates of parameters and should be done in a way such that every entry is consistent to some rule.

Rearranging equation (4)

$$\begin{aligned}
\beta_\ell &= \alpha_\ell \mathbf{1}_m + \lambda_\ell W_{m,\ell} \beta_\ell + \varepsilon(\beta_\ell) \\
\beta_\ell - \lambda_\ell W_{m,\ell} \beta_\ell &= \alpha_\ell \mathbf{1}_m + \varepsilon(\beta_\ell) \\
(I_m - \lambda_\ell W_{m,\ell}) \beta_\ell &= \alpha_\ell \mathbf{1}_m + \varepsilon(\beta_\ell) \\
\beta_\ell &= \alpha_\ell (I_m - \lambda_\ell W_{m,\ell})^{-1} \mathbf{1}_m + (I_m - \lambda_\ell W_{m,\ell})^{-1} \varepsilon(\beta_\ell)
\end{aligned}$$

where I_m is the $m \times m$ identity matrix. It can then be easily shown that the mean and the variance of β_ℓ is

$$\alpha_\ell (I_m - \lambda_\ell W_{m,\ell})^{-1} \mathbf{1}_m$$

and

$$(I_m - \lambda_\ell W_{m,\ell})^{-1} (I_m - \lambda_\ell W_{m,\ell}^T)^{-1} \sigma^2(\beta_\ell)$$

respectively, and due to linearity, we have

$$\beta_\ell \sim \mathcal{N}\left(\alpha_\ell (I_m - \lambda_\ell W_{m,\ell})^{-1} \mathbf{1}_m, (I_m - \lambda_\ell W_{m,\ell})^{-1} (I_m - \lambda_\ell W_{m,\ell}^T)^{-1} \sigma^2(\beta_\ell) \right)$$

for $\ell = 0, 1$. We can estimate λ_ℓ and α_ℓ by minimising the ordinary least squares equation in both λ_ℓ and α_ℓ. The ordinary least squares equation for β_ℓ is given by

$$\begin{aligned}
\left\{ \hat{\lambda}_\ell, \hat{\alpha}_\ell \right\} &= \underset{\lambda_\ell, \alpha_\ell}{\arg\min} \, ||\beta_\ell - \alpha_\ell \mathbf{1}_m - \lambda W_{m,\ell} \beta_\ell||_2^2 \\
&= \underset{\lambda_\ell, \alpha_\ell}{\arg\min} \left(\beta_\ell - \alpha_\ell \mathbf{1}_m - \lambda_\ell W_{m,\ell} \beta_\ell \right)^T \left(\beta_\ell - \alpha_\ell \mathbf{1}_m - \lambda_\ell W_{m,\ell} \beta_\ell \right).
\end{aligned} \tag{5}$$

We then minimise equation (5) with respect to λ_ℓ and α_ℓ by taking the derivatives, setting to zero and then solving simultaneously. This yields

$$\lambda_\ell = \frac{\beta_\ell^T (W_{m,\ell} + W_{m,\ell}^T)\beta_\ell - 2\alpha_\ell \beta_\ell^T W_{m,\ell}^T \mathbf{1}_m}{2\beta_\ell^T W_{m,\ell}^T W_{m,\ell}\beta_\ell}. \tag{6}$$

and

$$\alpha_\ell = \frac{1}{m}\mathbf{1}_m^T (I_m - \lambda_\ell W_{m,\ell})\beta_\ell \tag{7}$$

Then solving simultaneously gives estimators for λ_ℓ and α_ℓ as

$$\hat{\lambda}_\ell = \frac{\beta_\ell^T (W_{m,\ell} + W_{m,\ell}^T)\beta_\ell - \frac{2}{m}\beta_\ell^T W_{m,\ell}^T \mathbf{1}_m}{2\beta_\ell^T W_{m,\ell}^T W_{m,\ell}\beta_\ell - \frac{2}{m}\beta_\ell^T W_{m,\ell}^T \mathbf{1}_m \mathbf{1}_m^T W_{m,\ell}\beta_\ell} \tag{8}$$

and

$$\hat{\alpha}_\ell = \frac{2\mathbf{1}_m^T \beta_\ell \beta_\ell^T W_{m,\ell}^T W_{m,\ell}\beta_\ell - \beta_\ell^T (W_{m,\ell} + W_{m,\ell}^T)\beta_\ell \mathbf{1}_m^T W_{m,\ell}\beta_\ell}{2m\beta_\ell^T W_{m,\ell}^T W_{m,\ell}\beta_\ell - \beta_\ell^T W_{m,\ell}^T \mathbf{1}_m \mathbf{1}_m^T W_{m,\ell}\beta_\ell} \tag{9}$$

respectively. These estimators given in equations (8) and (9) have been shown through testing to be biased and inconsistent especially for λ when $||\beta||_2$ is large [10, 11], and is often accepted to be true [12], thus we should not use them. We can however use equations (6) and (7) in an iterative algorithm which updates at each step as following, where for this case $X = \mathbf{1}_m$,

Algorithm 1 Spatial Parameters Estimation (SPE) Algorithm

1: **procedure** SPE($\beta, W_{m,\ell}, X$)
2: Initialise $\lambda_\ell^{[0]} = 0$ and calculate for $i = 1, 2, \ldots$
3: **while** $|\lambda^{[i]} - \lambda^{[i-1]}| > \varepsilon$ **do**
4: $\alpha_\ell^{[i]} = \left(X^T X\right)^{-1} X^T \left(I - \lambda_\ell^{[i-1]} W_{m,\ell}\right)\beta_\ell$
5: $\lambda_\ell^{[i]} = \frac{\beta_\ell^T \left(W_{m,\ell}^T + W_{m,\ell}\right)\beta_\ell - \left(\beta_\ell^T W_{m,\ell}^T X \alpha_\ell^{[i]} + \alpha_\ell^{[i]^T} X^T W_{m,\ell}\beta_\ell\right)}{2\beta_\ell^T W_{m,\ell}^T W_{m,\ell}\beta_\ell}$
6: $i = i + 1$
7: **return** $\lambda^{[i]}$ and $\beta^{[i]}$

Once we have created these models with estimates for λ_ℓ and α_ℓ for $\ell = 0, 1$, we can then use these models to predict what the coefficients would be for a given location in Australia by spatial interpolation, i.e. let $\mathbf{x}_h$ be a location where there is no rain gauge and let $\mathbf{w}_{h,\ell} = \left(w_{h,1}^{[\ell]}, \ldots, w_{h,m}^{[\ell]}\right)^T$ be the spatial weight vector with each entry being a function of the distance metric from $\mathbf{x}_h$ to all other known coefficient locations (entries corresponding to locations with unknown coefficients are set to zero). Thus our estimate for the coefficients at location $\mathbf{x}_h$ are given by

$$\tilde{\beta}_\ell^{[h]} = \hat{\alpha}_\ell + \hat{\lambda}_\ell \mathbf{w}_{h,\ell}\beta_\ell \tag{10}$$

for $\ell = 0, 1$. From this we can then predict what the corresponding gauge measurement of location $\mathbf{x}_h$ for the j^{th} time interval would be from the corresponding satellite estimate by

$$\tilde{Y}_{hj}^{[g]} = \tilde{\beta}_0^{[h]} + \tilde{\beta}_1^{[h]} Y_{hj}^{[s]} \tag{11}$$

3 Results

The bureau of meteorology between January 2003 and August 2018 has had over 3000 rain gauge stations taking monthly precipitation measurements be in operation, 3368 of which we are using for this study ($m = 3368$). Figure 2 shows us the locations of the 3368 monthly rain gauge stations as well as the correlation between the precipitation measured by these gauges and what was estimated by the satellite imagery. As we can see, majority of these locations have reasonably high correlation between the gauge measurement and the satellite estimate, with 1921 out of the 3368 stations having a correlation factor greater than 0.7. The correlation factor also appears to be following a spatial trend, indicating that there may be a spatial trend in the relationship between gauge measurements and Satellite estimates. While gauge

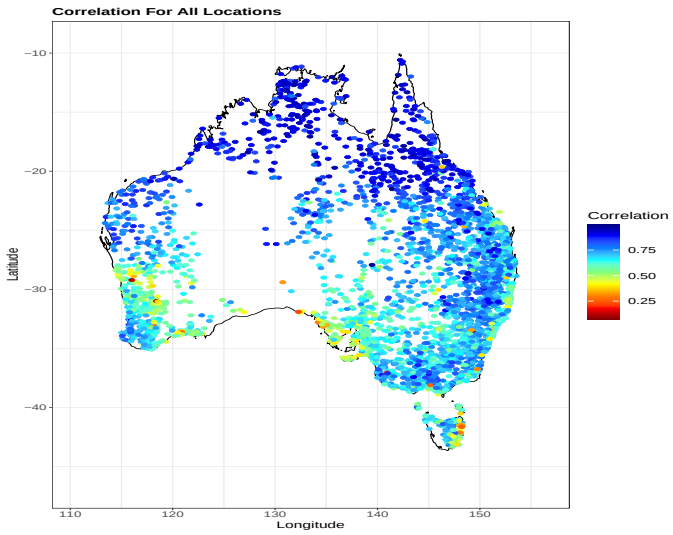

Fig. 2: Sample pearson correlation between the gauge measurement and the satellite estimate of the available 3368 Bureau of Meteorology rain gauge stations.

measurements and satellite estimates are highly positively correlated for the majority of locations we need to be able to justify that the relationship between the two is linear and also temporally stationary. Naturally we would assume that the

relationship between the gauge measurements and the satellite estimates is linear as in equation (1) as we would assume that the satellites give somewhat accurate estimates for the amount of precipitation, we have $\mathbb{E}\left[Y_{i,j}^{[g]}\right] = Y_{i,j}^{[s]}$. To justify this assumption, consider figure 3. We can see in the top left and top right plots the time

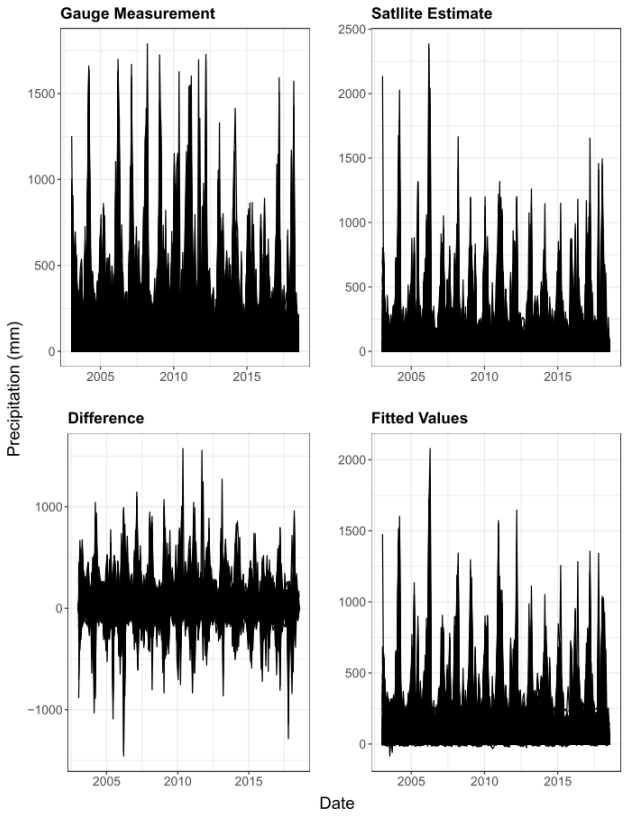

Fig. 3: Time series of all locations for gauge measurements (top left), satellite estimates (top right), difference of the gauge measurement and satellite estimate (bottom left) and linear regression fitted values (bottom right) in millimetres starting at January 2003 and ending at august 2018.

series for both the gauge measurements and satellite estimates respectively for all 3368 locations, where the satellite estimates are provided by the Japan Aerospace Exploration Agency (JAXA). In the bottom left plot of figure 3 we have the difference between the gauge measurements and the satellite estimates ($Y_{i,j}^{[g]} - Y_{i,j}^{[s]}$ for all $j = 1, \ldots, n_i$ and $i = 1, \ldots, m$) and we can see that on average the satellite imagery tends to overestimate the amount of precipitation that a gauge has measured. We can also see that there are points in the time series that there are very large differ-

ences between the gauge measurement and the satellite estimate in both directions. In the bottom right plot we fit the simple linear model to every location as described in equation (1) and recreate the time series in the top left plot using the fitted values of these models. We can see that it tends to recreate the rain gauge time series fairly well which gives evidence to the assumption of a linear relationship. However, the points where the gauge measurement and satellite estimate are significantly different suggests that there a significant outliers in the data. To justify the temporal stationarity assumption we can consider the difference between the gauge and satellite measurements at each location ($Y_{i,j}^{[g]} - Y_{i,j}^{[s]}$) as a time series then we can perform the augmented Dickey-Fuller test. We are testing the null hypothesis that there is a unit root present in the time series against the alternative that the time series is stationary [4]. The result of this test gives no locations with a p-value above 0.05 meaning the time series at each location are stationary. Meaning that the mean, variance and autocorrelation of the time series does not change with time [5], and thus we can justify using one model for each location with one set of coefficients that do not change with time.

In figure 4, we look at the total amount of precipitation recorded at each rain gauge station for the entire time that it was operational and compare it to the total estimates for that location over the same time period. We can see that whilst

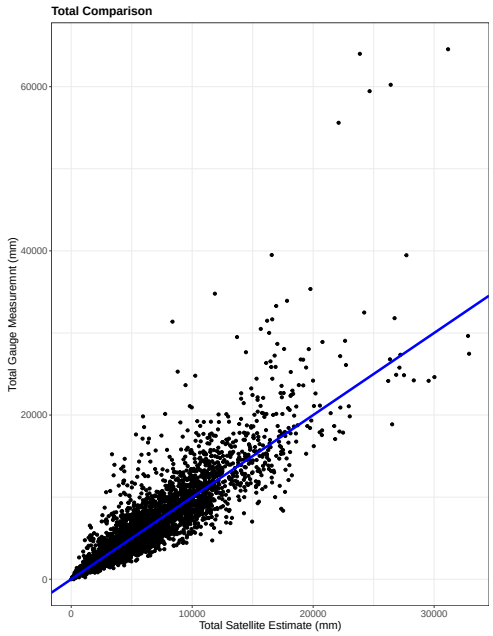

Fig. 4: Total gauge measurement of precipitation versus total satellite estimates of precipitation for each location over the same time interval compared to the line $y = x$.

the trend follows the red $y = x$ line quiet closely, giving more evidence to the assumption of a linear relationship, the variance in the difference between the total gauge measurements and the total satellite estimates increases as the measurements increases. If the satellite consistently only overestimates or only underestimates the gauge measurement in locations where there is a lot of precipitation, this problem will only be exacerbated when summing up the entire series. However, this issue could also be due to a single point in which the satellite estimate is significantly different to the gauge measurement as was seen in the bottom left plot of figure 3, which then skews the overall difference in the total comparison. In figure 5, we have the monthly gauge measurements plotted against their corresponding satellite estimates for rain gauge station 031030 (located on the eastern coast of northern Queensland, north of Cairns). The black $y = x$ line is what we would expect to see given that the satellite imagery gives an accurate estimate. We can see that this data has a significant outlier where the satellite estimates there to have been over 2200 millimetres of precipitation in a month compared to the gauge which only measured there to be just under 1000 millimetres of precipitation in that same month. Due to

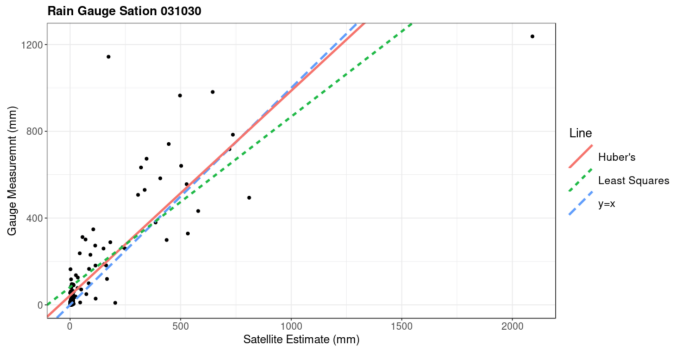

Fig. 5: Monthly gauge measurements of precipitation versus its corresponding monthly satellite estimates of precipitation for Station 031030, with the least squares regression line and the Huber's robust regression line.

the squared nature of the ordinary least squares in equation (2), these outliers can significantly influence the fit of the regression model (blue line) and thus causes the model to poorly capture the true trend of the relationship between the gauge measurement and the satellite estimate. We can reduce the influence of the significant outliers by using a robust regression method to model the relationship between the gauge measurements and the satellite estimates. Estimating parameters by robust regression requires us to minimise a different loss function as to what was shown in equation (2). We will be using Huber's loss to estimate $\beta^{[i]}$ in our robust regression model. Parameter estimation by Huber's loss is given by

$$\hat{\beta}_{\delta}^{[i]} = \underset{\beta_0^{[i]},\beta_1^{[i]}}{\arg\min} \begin{cases} \frac{1}{2}||\mathbf{Y}_i^{[g]} - \beta_0^{[i]}\mathbf{1}_{n_i} - \beta_1^{[i]}\mathbf{Y}_i^{[s]}||_2^2 & ,|Y_{ij}^{[g]} - \beta_0^{[i]} - \beta_1^{[i]}Y_{ij}^{[s]}| \leq \delta \\ ||\mathbf{Y}_i^{[g]} - \beta_0^{[i]}\mathbf{1}_{n_i} - \beta_1^{[i]}\mathbf{Y}_i^{[s]}||_1 - \frac{1}{2}\delta^2 & ,\text{otherwise} \end{cases} \tag{12}$$

where δ is a tuning parameter found by cross validation. In other words, we minimise the squared error for fitted values when they are within δ of the observed value and then minimise the absolute error for when fitted values are further than δ away from the observed value [8]. Figure 5 shows how the green robust regression line compares to the ordinary least squares regression line and we can see that by using Huber's loss, the significant outlier has much less influence on the fit of the regression line and which is a lot closer to the line $y = x$ than the ordinary least squares regression line.

In figure 6 we can see a clear trend in how the coefficients behave based on their location. As can be expected, higher values of the intercept coefficient, are

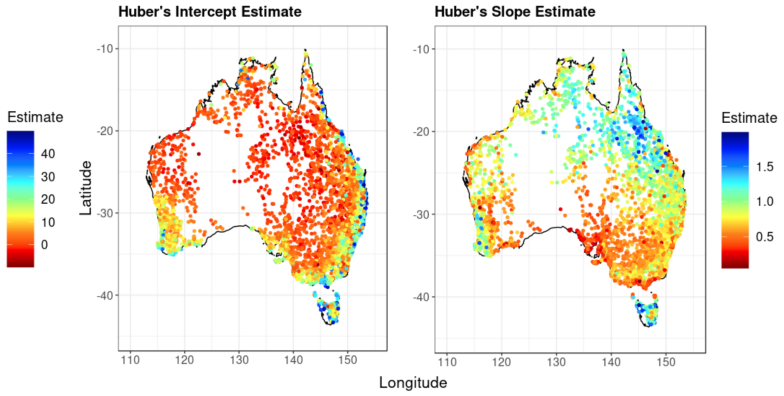

Fig. 6: Mappings of the Huber's regression coefficient estimates (intercept:left, slope:right) for the relationship between the gauge measurement and the satellite estimate for every location.

associated with lower values of the slope coefficient. There appears to be spatial trends in the relationship between the satellite estimate and the gauge measurement. Whilst a lot of coefficients are not giving values that we would expect with the intercept having a range of around -10 to 50 and the slope having a range of 0.05 to 1.8, this is due to the amount of noise that is present in data for some locations. While the noise is giving values for coefficients far off what we would expect to see for some locations, majority of locations have coefficients a lot closer to what we would expect to see. Figure 7 shows histograms of the coefficients estimated by Huber's loss in equation (12) we can see that the estimated values for the intercept are positively skewed with a mean of about 10 while the estimated values for the slope are quite symmetric with a mean of about 0.8.

To use a spatial linear model, we need to decide on what the weight matrices $W_{m,\ell}$ will be. As mentioned earlier, selection of a weight matrix is essential in creating

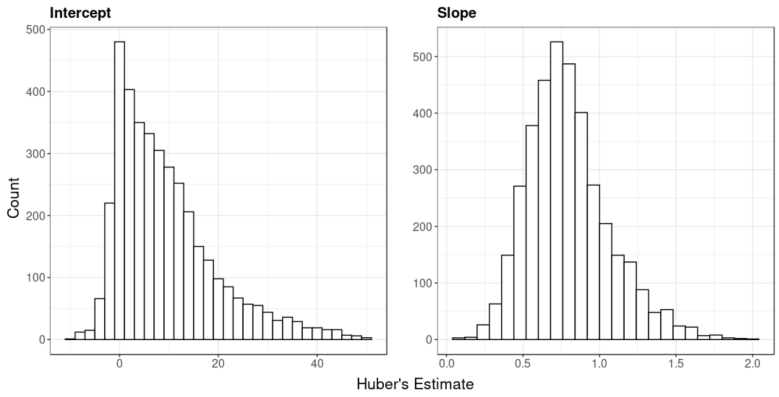

Fig. 7: Histograms of the estimated values of β_0 (left) and β_1 (right) by Huber's loss.

a good model. There are many different methods that could be used with different distance metrics. In this paper we use a mixture of k-nearest neighbours and inverse distance weighting (IDW) with an addition of the distance to the coast of the locations. That is,

$$w_{i,j} = \begin{cases} \frac{1/\left(d(\mathbf{x}_i,\mathbf{x}_j)^\gamma + dc(\mathbf{x}_i) + dc(\mathbf{x}_j)\right)}{\sum_{\mathbf{x}_t \in ne(\mathbf{x}_i)} 1/(d(\mathbf{x}_i,\mathbf{x}_t)^\gamma + dc(\mathbf{x}_i) + dc(\mathbf{x}_t))} & \text{, if } \mathbf{x}_j \in ne(\mathbf{x}_i) \\ 0 & \text{, otherwise} \end{cases} \tag{13}$$

where $ne(\mathbf{x}_i)$ is the neighbourhood of the location $\mathbf{x}_i$ which is determined by which other observed location are the k closest to it and $dc(\mathbf{x}_i)$ is the distance from the location $\mathbf{x}_i$ to the nearest coastal point. We use the distance to the coast as a factor in creating the weight matrices as we are working with rainfall data, the nature of the spatial relationships may change when a location is a further away from the coast, where there is significantly less rain. As the Surface of Australia lays on the surface of an approximate sphere, we should use a distance metric that considers the spherical nature of the domain. Thus we use a cosine distance metric, given by

$$d(\mathbf{x}_i,\mathbf{x}_j) = r\Delta\sigma$$

for

$$\Delta\sigma = \arccos\left(\sin x_{i,1}\sin x_{j,1} + \cos x_{i,1}\cos x_{j,1}\cos|x_{i,2} - x_{j,2}|\right)$$

with $r \approx 6371$ *km* (radius of the Earth). Whilst using the cosine distance metric doesn't significantly change the results compared to euclidean distance both in terms of the tuning parameters (k and γ) and also the end result, it is better to use a more accurate representation of the distance between locations.

Another well known neighbourhood defining method is the d-nearest neighbours method, where a neighbourhood for location $\mathbf{x}_i$ is defined to be the other locations

within a distance d of it. However, this method is not ideal for spatial processes where the observed locations are inconsistently placed as is gives a high amount of variance in the number of neighbours locations can have. For example, we can set our d to be a distance of 167 kilometres, which results in rain gauge station 13043 (South Karlamilyi National Park, Western Australia) having no neighbours, and rain gauge station 41103 (Toowoomba, Queensland) having 297 neighbours. The power coefficient γ is a non-negative value that represents the smoothness of the weight matrix. By cross-validation we can get the optimal values for creating the weight matrices for the intercept and slope as $(k=7,\gamma=0.92)$ and $(k=8,\gamma=0.79)$ respectively.

$$\begin{aligned}\hat{\beta}_0 &= 0.122\times\mathbf{1}_m + 0.983\times W_{m,0}\beta_0\\ \hat{\beta}_1 &= 0.021\times\mathbf{1}_m + 0.972\times W_{m,1}\beta_1\end{aligned} \tag{14}$$

respectively. The estimates for λ_0 and λ_1 as 0.983 and 0.972 respectively have associated likelihood ratio p-values that are significantly small ($< 10^{-100}$) which indicate that there is a high degree of spatial dependency. Figure 8 shows us how the

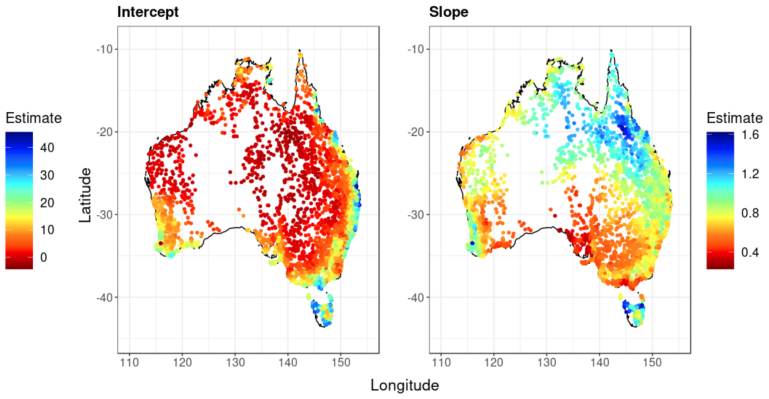

Fig. 8: Spatially modelled fitted values for the intercept (left) and slope (right) with estimation done by the SPE algorithm.

equations in (14) model the coefficients and as we can see, our models fit the data quite well as they appear to have captured the spatial trends present in the data. We can also see that locations with their coefficients significantly different to the surrounding locations coefficients' are not estimated as well by the model.

The fitted values that are not estimated as well as their surrounding coefficients may be able to be modelled better with the inclusion of a confounding factor. Thus far we have only looked at longitude and latitude contributing to how our coefficients are estimated by model (4), but it is possible that the elevation dimension is significant in modelling a coefficient's behaviour. Elevation could help us to explain a coefficient's behaviour as the satellite's image in which the precipitation is

estimated from is taken from above, and rain gauge stations that are at a higher elevation will be closer to the cloud and the satellite. Whilst the difference in elevation is marginal compared the altitude of the satellite, it is not marginal compared the altitude of the clouds. The elevation range of the rain gauge stations is $1,868$ meters, with multiple at sea level and the highest located in Kosciuszko National Park, New South Wales. We can see in figure 9 an elevation mapping of Australia with locations of interest being labelled. Precipitation causing clouds can occur anywhere between ground level and 6000 metres [7], therefore the altitude of the rain gauge station may help us better model the coefficients' behaviour. This new model with

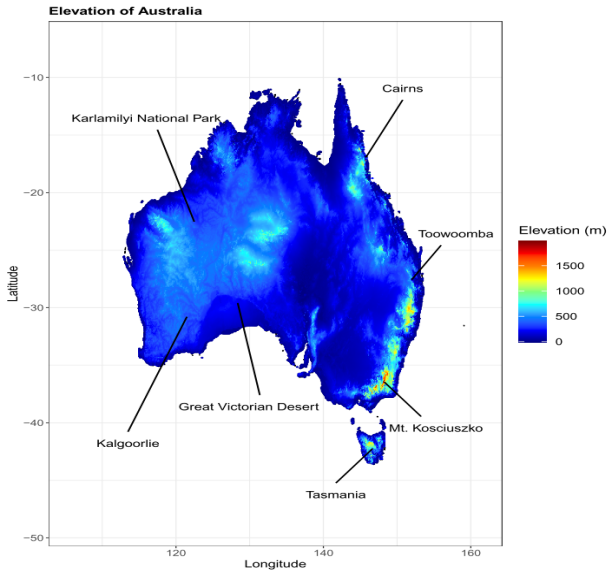

Fig. 9: Map of Australia showing the elevation (m).

the elevation confounding factor included can be expressed as

$$\beta_\ell = \alpha_\ell \mathbf{1}_m + \eta_\ell \mathbf{e} + \lambda_\ell W_{m,\ell} \beta_\ell + \varepsilon(\beta_\ell) \tag{15}$$

for $\ell = 0, 1$, with $\mathbf{e}$ being an $m \times 1$ vector where e_i is the elevation of the rain gauge station at location $\mathbf{x}_i$ for $i = 1, \ldots, m$. We can again use the SPE algorithm to find estimates for the unknown parameters with $X = (\mathbf{1}_m, \mathbf{e})$. We find here that the optimal values for defining the weight matrices for the intercept is the same, but the slope is now $(k = 8, \gamma = 0.8)$. For the spatial linear model of the intercept, the coefficient η_0 of the elevation confounding factor is estimated to be 0.0016 with an associated p-value of 5×10^{-5}, lowering the residual error from 5.429 to 5.417. For the slope coefficient spatial model, the elevation parameter is significant with the coefficient η_1 being estimated to be 6.3×10^{-5} with an associated p-value of 4.6×10^{-7}, lowering the residual error from 0.1502 to 0.1496. Therefore, the equations for the

intercept and slope models are now given by

$$\begin{aligned}\hat{\beta}_0 &= -0.310 \times \mathbf{1}_m + 0.0016 \times \mathbf{e} + 0.986 \times W_{m,0}\beta_0 \\ \hat{\beta}_1 &= 0.012 \times \mathbf{1}_m + 6.3 \times 10^{-5} \times \mathbf{e} + 0.964 \times W_{m,1}\beta_1\end{aligned}. \tag{16}$$

Now that we have estimates for the parameters α_ℓ, η_ℓ and λ_ℓ, we can recreate the

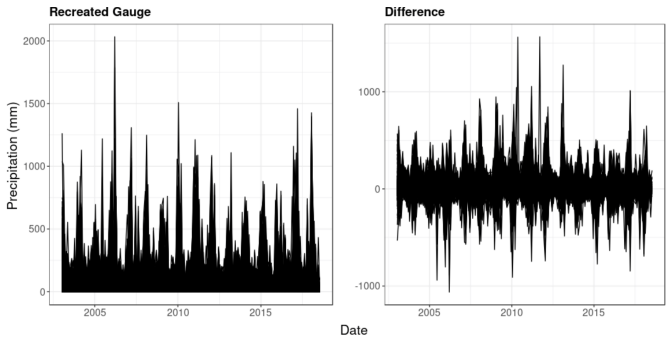

Fig. 10: Recreated time series using the fitted values from equations (16) for all observed locations (left) and the difference between the gauge measurements and the recreated gauge measurements (right).

time series of observations and compare this to the original gauge measurements time series as in figure 3. In the left plot of figure 10 we recreated the time series using the fitted values from the models in equation (16) for all observed locations over all time intervals. We can see that this method does yield a similar result to what the simple linear model used for the bottom right plot of figure 3 indicating that the spatial linear model does a good job of recreating the coefficients estimated by the linear model in equation (1). We can also see in the right plot of figure 10 that difference between the gauge measurements and the recreated gauge measurements is for the most part, around zero. There are many points where the gauge measurements and recreated gauge measurements are significantly different, however we must remember that we are using Huber's estimator as a loss function instead of the ordinary least squares and therefore less weight is given to 'outliers' so these points are not estimated as well by the model.

We can proceed to predict how the intercept and the slope would behave in regions where there are very few or even no rain gauges. We generate 50000 grid points over all of Australia. We then define the weight matrix $\mathbf{w}_{h,\ell}$ for each of the new points from their neighbourhoods as defined in equation (13) with $(k = 7, \gamma = 0.92)$ and $(k = 8, \gamma = 0.8)$ for the intercept and slope respectively. Recall that $\mathbf{w}_{h,\ell}$ will only depend on the existing rain gauge locations. We then use equation (10) with the addition of the elevation confounding factor

$$\tilde{\beta}_\ell^{[h]} = \hat{\alpha}_\ell + \hat{\eta}_\ell e_h + \hat{\lambda}_\ell \mathbf{w}_{h,\ell}\beta_\ell$$

to estimate the intercept and slope for location $\mathbf{x}_h$, where e_h is the elevation at $\mathbf{x}_h$ for all new locations ($h = m+1,\ldots,m+50000$). Note that due to the sparsity of existing rain gauges in some areas many of these new points will be defined to have the same neighbourhoods which results in their coefficients to be predicted as the same. Figure 11 shows how the models would predict these new points and as we can see, these added points follow the spatial dependency we had expected to see. Following from here, the gauge measurement time series at those new locations can be estimated. For example, consider map coordinates longitude 126.284 and latitude -29.068, located in Western Australia just east of the Great Victoria Desert Nature Reserve and about 500 kilometres from Kalgoorli at an elevation of 232 metres. The models in equation (16) give the estimated intercept and slope for this location as 3.42 and 0.46 respectively. The JAXA satellite estimate for the month of August 2018 is 10.3mm, substituting 10.3 into equation (11) gives 8.2mm as the gauge replicate for that location for month of August 2018. There are other loss functions

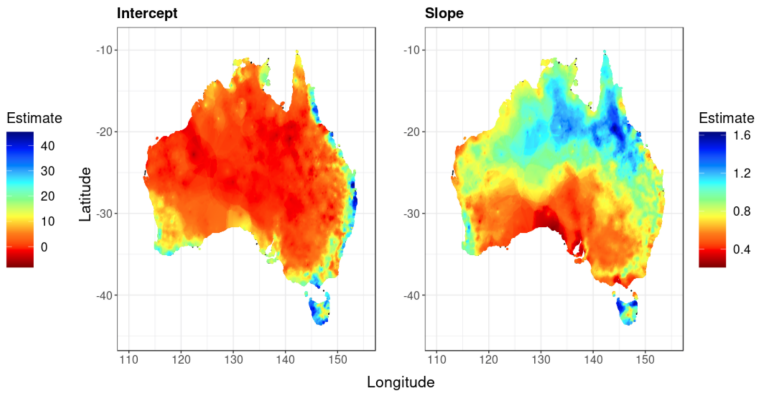

Fig. 11: Spatially Predicted value plots for the intercept (left) and slope (right) of the Huber's regression coefficients of the relationship of the gauge measurement against the satellite estimate of precipitation using the SPE algorithm.

that we could use besides the ordinary least squares to develop an algorithm for estimating the spatial parameters of a model as in equations (4) and (15). We could use a robust regression but as can be seen in by the histograms in figure 7 there are no significant outliers that will skew the estimation of the parameters by having too much leverage, thus there is no need to use robust regression in this case. However, at the moment we are estimating the spatial parameters of the intercept and slope separately when the intercept and slope are obviously dependent. We can also use the observed values from the gauge measurements and the satellite estimates in our new minimisation

$$\sum_{i=1}^{m} \left|\left| \mathbf{Y}_i^{[g]} - \mathbf{1}_{n_i} \left(\mathbf{x}_i^T \alpha_0 + \lambda_0 \mathbf{w}_{0i}^T \beta_0 \right) - \mathbf{Y}_i^{[s]} \left(\mathbf{x}_i^T \alpha_1 + \lambda_1 \mathbf{w}_{1i}^T \beta_1 \right) \right|\right|_2^2 \tag{17}$$

where $\alpha_\ell = (\alpha_\ell, \eta_\ell)$. By taking the derivative of equation (17) with respect to the spatial parameters $\lambda_0, \lambda_1, \alpha_0$ and α_1, setting to zero and rearranging, we can obtain 4 equations for each of the parameters which all depend on the other parameters. We can then create another algorithm similar to the SPE algorithm that can give estimates for the spatial parameters. This algorithm gives parameter estimates as the following

$$\begin{array}{lll} \hat{\lambda}_0 = 1.17 & \hat{\alpha}_0 = 0.97 & \hat{\eta}_0 = -0.02 \\ \hat{\lambda}_1 = 0.9 & \hat{\alpha}_1 = 0.06 & \hat{\eta}_1 = 0.0003 \end{array}.$$

Note that in some literature there exists the restriction of $|\lambda_\ell| < 1$ [1]. We can see in figure 12 how this new loss function in equation (17), where $\mathbf{w}_{0i}$ and $\mathbf{w}_{1i}$ are defined the same as above, alters the prediction of the coefficients, especially in locations where the number of gauge locations is sparse. Interestingly, the range of the

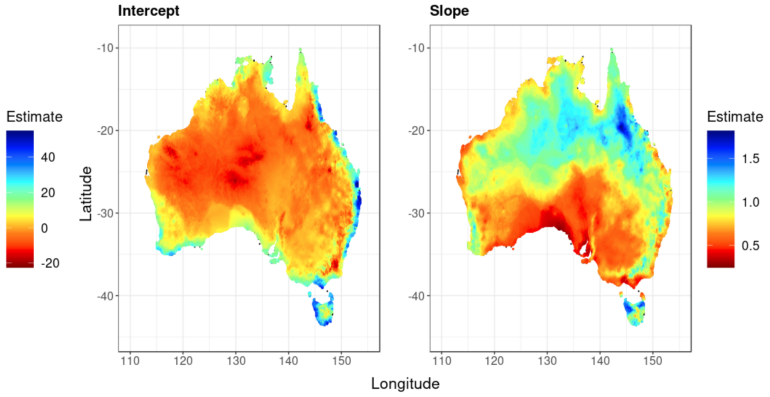

Fig. 12: Spatially Predicted value plots for the intercept (left) and slope (right) of the Huber's regression coefficients of the relationship of the gauge measurement against the satellite estimate of precipitation using the loss function in equation (17).

predicted slope values has slightly decreased where as the range for the predicted intercept values has slightly increased. Again we can give an estimate for what the gauge measurement would read in a location where there are no rain gauges, in the same location as used above, the new estimates for the intercept and slope are now given as -0.487 and 0.538 respectively which gives the recreated gauge measurement as 5.05, significantly lower than the previously estimated gauge measurement.

4 Conclusion

There is clear evidence that there is some difference between the gauge measurements and the satellite estimates. For the majority of locations, the satellite estimates on average that slightly more precipitation has occurred than what the gauge has measured for the same time interval. There is also clear evidence of spatial dependency for the relationship between the gauge measurement and satellite estimate.

The spatial linear model appears to be able to model the spatial dependencies of the robust regression coefficients well. A problem with attempting to predict how the intercept and slope coefficients would behave for locations where there are no rain gauges (as shown in figure 11), is that there is potential for areas to have very localised spatial behaviour. This localised spatial behaviour of the intercept and slope coefficients may not have been captured as the neighbours for these locations are far away and may be behaving differently based on their own localised spatial behaviour. We can see in the plots of figure 6, in locations such as Tasmania and north-east Australia that there appears to be very localised behaviour.

In comparing each of the spatial models used, they each have their own pros and cons. The first model in equation (4) while good as it easy to compute and easy to interpret, the model is probably overly simple as it does not use any confounding factors, the estimation is for the coefficients is done independently of each other and it does not depend on the gauge measurements or the satellite estimates. The second model in equation (15) is also easy to compute and it uses elevation which was found to be statistically significant for both the intercept and the slope. However, similarly to the first model, it is not good that this model also does estimation of the coefficients independently and does not depend of the gauge measurements of the satellite estimates. The last model using the loss function in equation (17) performs better than the other two models in fitting the known gauge measurements due to the parameters being estimated dependently and including the gauge measurements and satellite estimates. However, due to the parameters being estimated dependently there is a larger residual error of the coefficients and it more computationally complex than the other two models.

The amount of noise with the readings does present a problem in itself. The large amount of noise makes for the intercept's coefficient in some locations to be so large, that even if there was no precipitation estimated by the satellite images for the month, the model could give that a rain gauge would have recorded 50mm. One way to adjust for this would be to have each observation be a measured/estimated over a greater time interval such as 3 months or even a year instead of only a month. An issue with increasing the length of the time interval is that there are some rain gauge stations that were only in operation for a couple years or even less and increasing the length of the time interval would decrease what little amount of observations they had making for a less accurate estimation of the coefficients. Another approach to reduce the influence of the amount of noise on the fit of the model is to use an even more robust loss function to estimate the intercept and slope coefficients.

There is potential for more confounding factors to be included in the spatial linear models that may help explain the spatial relationship between the gauge mea-

surements and the satellite estimates. However, we must keep in mind that with spatial modelling that including too many explanatory variables can result in spatial over-fitting where the explanatory variables explain the spatial dependencies in the process, i.e. as the number of explanatory variables increases, $|\hat{\lambda}_\ell|$ decreases [14]. While in this paper we have modelled the relationship between the gauge measurements and the satellite estimates for each location by a strictly linear relationship, this may not be the case, the relationship may be more complex and to properly model this relationship we may need to consider transformations of the variables. We could also explore more complicated weight functions, that could include more spatially descriptive factors, or have the distance to the coast factor be included in a different way.

The loss function in equation (17), which has parameter estimation depending of the intercept and the slope depending on each other and also the gauge measurements and the satellite estimates, gives interesting results for the coefficient predicted as in figure 12. There are alterations that could be made to this loss function, such as the minimisations at each location are given weights. We could do this in many ways such as the weights being dependent on how many observations are at a given location, giving more weight to locations with more observations, or we could define weights depending on how isolated the observed location is.

References

1. Arbia, G.: A primer for spatial econometrics: with applications in R. Springer (2014)
2. Beenstock, M., Felsenstein, D., et al.: The Econometric Analysis of Non-Stationary Spatial Panel Data. Springer (2019)
3. Freedman, D.A.: Statistical models: theory and practice. cambridge university press (2009)
4. Fuller, W.A.: Introduction to statistical time series, vol. 428. John Wiley & Sons (2009)
5. Hamilton, J.D.: Time series analysis, vol. 2. Princeton university press Princeton, NJ (1994)
6. Herrera, M., Mur, J., Ruiz Marin, M.: Selecting the most adequate spatial weighting matrix: A study on criteria. Tech. rep., University Library of Munich, Germany (2012)
7. Houze Jr, R.A.: Cloud dynamics, vol. 104. Academic press (2014)
8. Huber, P.J., et al.: Robust estimation of a location parameter. The annals of mathematical statistics **35**(1), 73–101 (1964)
9. Kachi, M., Kubota, T., Ushio, T., Shige, S., Kida, S., Aonashi, K., Okamoto, K., Oki, R.: Development and utilization of"jaxa global rainfall watch"system based on combined microwave and infrared radiometers aboard satellites. IEEJ Transactions on Fundamentals and Materials **131**, 729–737 (2011)
10. Lee, L.F.: Asymptotic distributions of quasi-maximum likelihood estimators for spatial autoregressive models. Econometrica **72**(6), 1899–1925 (2004)
11. Lee, L.f., Yu, J.: Estimation of spatial autoregressive panel data models with fixed effects. Journal of Econometrics **154**(2), 165–185 (2010)
12. Li, H., Calder, C.A., Cressie, N.: Beyond Moran's I: testing for spatial dependence based on the spatial autoregressive model. Geographical Analysis **39**(4), 357–375 (2007)
13. Ord, K.: Estimation methods for models of spatial interaction. Journal of the American Statistical Association **70**(349), 120–126 (1975)
14. Schabenberger, O., Gotway, C.A.: Statistical methods for spatial data analysis. Chapman and Hall/CRC (2017)