SPATIAL ESTIMATION OF AVERAGE DAILY PRECIPITATION USING MULTIPLE LINEAR REGRESSION BY USING TOPOGRAPHIC AND WIND SPEED VARIABLES IN TROPICAL CLIMATE

Mohd Talha ANEES[1]*, Khiruddin ABDULLAH[1], M. N. M. NAWAWI[1],
Nik Norulaini Nik AB RAHMAN[2], Abd. Rahni MT. PIAH[3], M. I. SYAKIR[4],
Mohammad Muqtada ALI KHAN[5], Abdul Kadir MOHD. OMAR[4]

[1]*School of Physics, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia*
[2]*School of Distance Education, Universiti Sains Malaysia, 11800 Minden, Penang, Malaysia*
[3]*Postgraduate School, DRB-HICOM University of Automotive Malaysia, 26607 Pekan, Pahang, Malaysia*
[4]*School of Industrial Technology, University Sains Malaysia, 11800, Minden, Penang, Malaysia*
[5]*Faculty of Earth Science, Universiti Malaysia Kelantan, Jeli Campus, Locked Bag No. 100,*
17600 Jeli, Kelantan, Malaysia

Abstract. Complex topography and wind characteristics play important roles in rising air masses and in daily spatial distribution of the precipitations in complex region. As a result, its spatial discontinuity and behaviour in complex areas can affect the spatial distribution of precipitation. In this work, a two-fold concept was used to consider both spatial discontinuity and topographic and wind speed in average daily spatial precipitation estimation using Inverse Distance Weighting (IDW) and Multiple Linear Regression (MLR) in tropical climates. First, wet and dry days were identified by the two methods. Then the two models based on MLR (Model 1 and Model 2) were applied on wet days to estimate the precipitation using selected predictor variables. The models were applied for month wise, season wise and year wise daily averages separately during the study period. The study reveals that, Model 1 has been found to be the best in terms of categorical statistics, R^2 values, bias and special distribution patterns. However, it was found that sets of different predictor variables dominates in different months, seasons and years. Furthermore, necessities of other data for further enhancement of the results were suggested.

Keywords: inverse distance weighting, interpolation, multiple linear regression, precipitation occurrence, spatial distribution, wind speed.

Introduction

Uncertainties, especially input variables, in watershed hydrological modeling are great concern for researchers (Carpenter & Georgakakos, 2004). Precipitation is one of the most basic meteorological input variables in hydrologic simulation to understand either flood risk or soil loss estimation at within a watershed (Mikoš, Jošt, & Petkovšek, 2006; Johnson, Hutchinson, The, Beesley, & Green, 2016). In a complex topography, the spatial and temporal behaviour of precipitation are generally influenced by the variations in relief, easting, northing, slope and strong wind (Allamano, Claps, Laio, & Thea, 2009; Tao, 2009; Qing, Zhu-Guo, & Liang, 2011; Hwang, Clark, Rajagopalan, & Leavesley, 2012). For accurate characterization of spatial precipitation, particularly in complex relief regions, dense rain gauges network are needed which is very difficult in terms of installation and costs (Mair & Fares, 2010). Various interpolation methods have been used by researchers to solve this problem but their accuracies vary in different climates. Therefore, the choice of an interpolation method requires the understanding of the spatial variability of precipitation and the sources of uncertainty (Tao, 2009).

Several simple methods such as simple averaging, Thiessen polygons, isohyetal and Inverse Distance Weighting (IDW) have been used so far as traditional methods in spatial estimation of precipitation (Thiessen, 1911; Shepard, 1968; Tabios & Salas, 1985; McCuen, 1989). But these methods do not include any physical predictor variables. As an alternative, complex statistical methods such

*Corresponding author. E-mail: talhaanees_alg@yahoo.in

as Simple and Multiple linear regression (SLR and MLR) and locally weighted polynomial (LWP) are widely used models which can correlate precipitation with physical predictor variables (Rajagopalan & Lall, 1998; Goovaerts, 2000; Kurtzman, Navon, & Morin, 2009; Hwang et al., 2012). Geostatistical methods such as kriging and co-kriging are other widely used methods for spatial distribution of precipitation. Some other methods having fewer advantages over traditional methods have been reported by (Goovaerts 2000; Drogue, Humbert, Deraisme, Mahr, & Freslon, 2002; Buytaert, Celleri, Willems, De Bievre, & Wyseure, 2006). However, geostatistical methods are used generally for monthly and annual data (Diodato, 2005; Mair & Fares, 2010; Gonga-Saholiariliva, Neppel, Chevallier, Delclaux, & Savéan, 2016) because these methods are not easy to apply on daily estimation of precipitation in a complex topography (Ly, Charles, & Degre, 2011; Castro, Gironás, & Fernández, 2014). Furthermore, the accuracy of different methods varies from region to region (Hwang et al., 2012).

Interpolation methods for spatial distribution of precipitation is restricted as there is an uncertainty called discontinuity in daily precipitation which affects spatial distribution of precipitation in complex topography. Previous studies used regression models such as Precipitation-elevation Regression on Independent Slope Model (PRISM) and Auto-Search Orographic and Atmospheric Effects Detrended Kriging (ASOADeK) (Daly, Neilson, & Phillips, 1994; Guan, Wilson, & Makhnin, 2005; Xie et al., 2007) by including orographic and meteorological predictor variables. Few studies have include wind speed as a predictor variables (Johansson & Chen, 2003; Allamano et al., 2009) but none of the studies considered spatial discontinues of precipitation.

Some of recent studies include discontinuity of precipitation (called phase estimation or occurrence/non-occurrence or wet/dry) and successfully estimated daily spatial precipitation by including different predictor variables (Seo, 1998; Hewitson & Crane, 2005; Hwang et al., 2012; Castro et al., 2014). Hewitson and Crane (2005) used conditional interpolation method for phase estimation as a function of the synoptic state in sub-tropical climate. Their method of estimation was based on the ability to reproduce the frequency of events, rather than the errors in the magnitude of the estimations as discussed by Castro et al. (2014). While Hwang et al. (2012) used daily logistic regressions to classify occurrence/non-occurrence based on monthly threshold and then applied four interpolation methods (IDW, MLR, LWP and Climatological MLR) on wet days by including three predictor variables (northing, easting and elevation). Castro et al. (2014) also estimated phase by IDW based method and then used IDW and SLR methods on wet days by including elevation and slope as a predictor variable in the climate between Mediterranean and mildly humid. They classified slope orientation either on windward or on leeward side with respect to the prevailing wind direction which gives better results than IDW

and SLR. Hwang et al. (2012) and Castro et al. (2014) both used elevation as the main predictor variable because of the importance of orographic barriers in uplifting air masses transported by wind which generate significant precipitation at high relief. However, maximum precipitation at highest point might not be necessary (Daly et al., 1994). Furthermore, the authors highlighted the necessity of multiple linear regression along with other predictor variables such as wind characteristics, relative humidity and distance from shoreline to enhance the results in precipitation estimation.

None of the studies related to discontinuity of precipitation have been reported in tropical climate. The tropical climate is important in this regard because of the dominancy of precipitation throughout the year (Anees et al., 2017). Therefore, in terms of phase estimation, the study used two methods (one is previously used IDW based phase estimation method and second our proposed simple threshold method) to compare their performance because of monthly dynamic nature of precipitation in tropical climate. The reason behind the selection of the previously used phase estimation method is to check its performance in tropical climate. Furthermore, the study includes northing, easting, elevation, slope and wind speed as predictor variables through multiple linear regression because of the necessities highlighted above in tropical climates.

The main foci of this study are (i) the methodology development for phase estimation (ii) to improve average daily precipitation interpolation technique and (iii) to know the effect of the predictor variables on estimation of average daily precipitation (month wise, season wise and year wise) in tropical climate. Hence, the aims of the study are: (i) the phase estimation of precipitation by using two methods to find wet and dry days and (ii) to use the MLR on wet days (obtained from both the methods of phase estimation) and to compare the performance of the two methods and (iii) to check the proposed interpolation methodology in daily precipitation estimation. The two algorithms incorporated the selected predictor variables to estimate average daily precipitation for each grid cells.

1. Materials and methods

1.1. Study area and data

The study was conducted in Kelantan state, north eastern part of Peninsular Malaysia between latitudes 4° 33′ and 6°14′ North, and longitudes 101° 19′ and 102° 39′ East with an area of approximately 15 000 km². The northern part is bounded by South China Sea and higher elevations are on south, south-east and south-west side (Figure 1). The elevation ranges from 0 to the highest point of 2.187 m named as Mt. Tahan which is situated on Mount Korbu (the second highest mountain of Peninsular Malaysia). The length and breadth the catchment is 187 km and 148 km respectively. The main river in the study area is Sungai Kelantan which divides after approximately 107 km from the mouth of the basin into Sungai Galas and Sungai Lebir.

The climate is tropical, humid with average temperature ranges from 20 °C to 30 °C with annual variation of 2 to 3 °C and average relative humidity of 82%. Minimum and maximum relative humidity usually occurs in March and November respectively. The period from November to January receives maximum rainfall, while June and July are the driest months. According to Malaysian Meteorological Department, the wind strength is generally light and variable, however, there are yearly periodic changes in the wind flow patterns and based on this, the season is divided into four, namely, the southwest monsoon, northeast monsoon and two shorter periods of inter-monsoon seasons. The state is affected by the North-eastern monsoon wind accompanied by heavy precipitation beginning early November and ends in March. The northeast monsoon occur due to the development of high pressure area at Asian continent (because of lowering temperature in winter season) and the development of low pressure area at Australian continent (because of high temperature in summer season) which results the movement of wind from the high pressure area of Asia to the low pressure area of Australia. The easterly or north-easterly winds of 10 to 20 knots can sometimes reach 30 knots or more as a result of the strong surges of cold air from the north.

Masseran and Razali (2016) mentioned that the winds blow across the South China Sea brings heavy rain, especially to the eastern and southern coasts of Peninsular Malaysia as well as the central part of the Titiwangsa Ranges. The study area is more affected by northeast monsoon which causes flooding, especially in November and December in the east coast states. Average annual rainfall of the area is 3017.84 mm while average daily annual wind speed is 1.50 m/s. Generally about 7 hours of sunshine is received by the coastal areas and about 6 hours of sunshine is received by the rest of the area. Therefore, evaporation rate is high near the coastal areas (4 to 5 mm/day) as compared to the high land areas (2.5 mm/day).

Daily rainfall data from 56 rain gauge stations were obtained from the Department of Irrigation and Drainage, Malaysia. Out of the 56 stations, 25 are relatively new which have data recorded from year 2005 onwards. Daily records, measured from current day at 12:00 am up to the next day at 12:00 am, for each station from January 2005 to December 2015 were selected. The daily mean precipitation for this period was 7.27 mm/d. Figure 2 shows the overall precipitation for the selected period with Coefficient of determination $(R^2) = 0.64$.
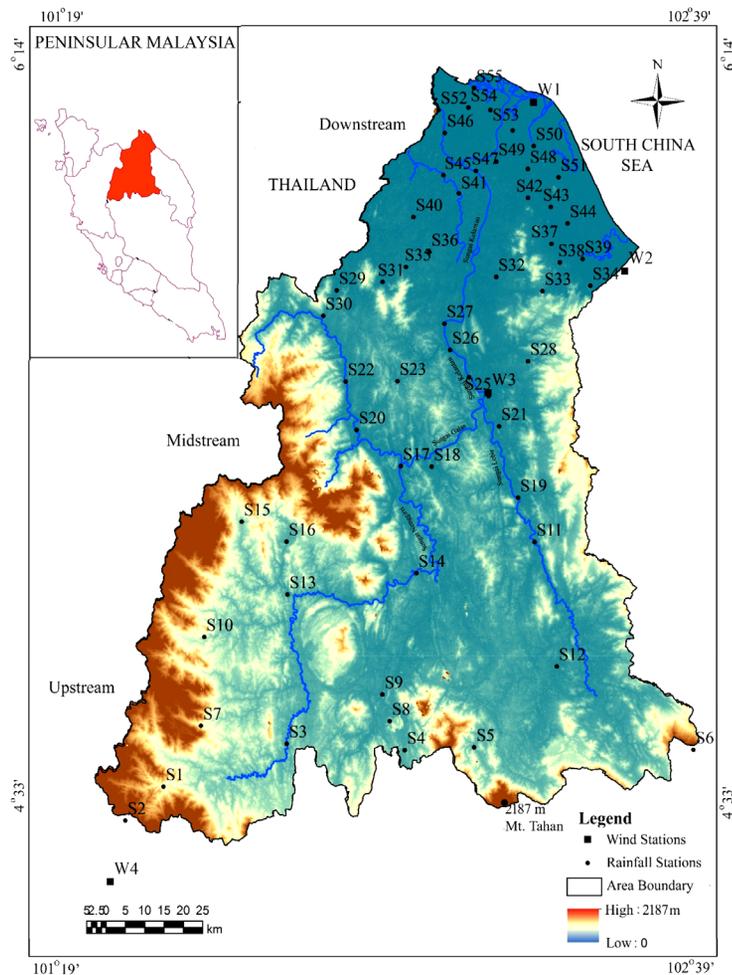
Figure 1. Study area with rain gauges, wind speed stations and elevation
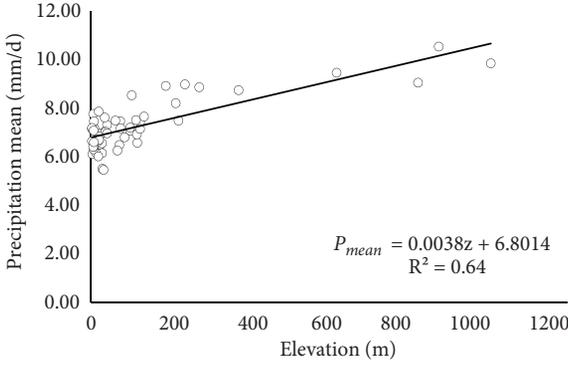
Figure 2. Mean annual precipitation (P_{mean}) for each rain gauge vs. elevation

The R^2 for precipitation with northing, easting, slope and wind speed are 0.68, 0.32, 0.08 and 0.42 respectively (Figure S1).

The mosaic of ASTER Global Digital Elevation Model (GDEM) of 30 m resolution was downloaded (http://gdem.ersdac.jspacesystems.or.jp/download.jp) to cover the study area in which the slope was calculated by Geographic Information System (GIS). The area has only three wind speed stations and one at a nearby area to the south (Figure 1) provided by Malaysian Meteorological Department. Wind speed at each stations were estimated by traditional Inverse Distance Weighted (IDW) interpolation method. Wind direction is not included in this study because the area has generally three dominant directions in a day. Therefore, taking daily average of wind direction could have affected the results.

1.2. Methodology

The proposed methodology is based on a two-fold approach. First, to identify wet and dry days by two methods in each grid cell (30 m resolution) of the surrounding rain gauges. Second, to apply MLR on wet days by including the predictor variables in order to assess the influence of these variables on the magnitude and distribution of precipitations in the area. The proposed methodology is shown in Figure 3.

1.2.1. Phase estimation method 1

In this phase estimation, the precipitation for each grid cell is estimated by using the traditional IDW method as used by Hwang et al. (2012):

$$P_{est} = \sum_{i=1}^{n} w_i p_i ; \tag{1}$$

$$w_i = \frac{\dfrac{1}{d_i^k}}{\sum_{j=i}^{n} \dfrac{1}{d_k^j}} , \tag{2}$$

where n is the number of observations within radius of influence (R_{inf}); d_i is the distance between a target and i^{th} observations; d_j is the distance between the target and

each of jth observations and $1 \leq k \leq \infty$. R_{inf} was selected so as to get a minimum number of three rain gauges. The first method of phase estimation is simple threshold (Method 1) of daily data in which $P_{est} = 0$ for dry days and $P_{est} > 0$ for wet days referred to as Data 1.

1.2.2. Phase estimation method 2

Second, IDW based phase estimation method (Method 2) of daily data, proposed by Castro et al. (2014), which states that the precipitation occurrence (ω_c^j) is estimated as a distance weighted average of the occurrence in the n surrounding rain gauges which is given by:

$$\omega_c^j = \frac{\sum_{i=1}^{n} W_i \varphi_i^j}{\sum_{i=1}^{n} W_i} ; \tag{3}$$

$$\varphi_i^j = \begin{cases} 0, & \text{if } P_i^j = 0 \\ 1, & \text{if } P_i^j > 0 \end{cases} ; \tag{4}$$

$$W_i = \begin{cases} 0, & \text{if } d_i > R_{inf} \\ \dfrac{\left(\dfrac{1}{d_i}\right)^b}{\sum_{i=1}^{n} \left(\dfrac{1}{d_i}\right)^b}, & \text{if } d_i \leq R_{inf} \end{cases} , \tag{5}$$

where c is grid cell; φ_i^j is a variable that incorporates the phase according to the daily precipitation record P_i of each rain gauge i, for each day j. W_i is the weighting factor for distance d_i and gauge i, and b is a parameter controlling the weighting factor. The phase estimation was carried out for each grid-cell and day of the study period on the basis of ω_c^j value. If $\omega_c^j < 0.5$, the grid cell c on day j is classified as dry while $\omega_c^j \geq 0.5$ as wet. The wet days measured by second phase estimation method are referred to as Data 2.

1.2.3. Multiple linear regression (MLR) based models

After the selection of Data 1 and Data 2 (based on whether method 1 or method 2 is to be used), the average of each day over the whole study period is taken which will be a dependent variable in MLR while independent variables will be the predictor variables. The model can be formulated as:

$$P_{est} = \beta_0 + \beta_1 x + \beta_2 y + \beta_3 z + \beta_4 s + \beta_5 w, \tag{6}$$

where x, y, z, s and w are predictor variables of northing (m), easting (m), elevation (m), slope (degree) and wind speed (m/s) respectively. β_0, β_1, β_2, β_3, β_4, β_5 are the regression coefficients which are estimated by minimizing the least squared errors. The interpolation of Data 1 is referred to as first model named PMLRW 1 (Phase Multiple Linear Regression and Variable Selection 1, hereafter it will be referred to as Model 1) whereas the interpolation of Data 2 is referred as second model namely PMLRW 2 (Phase Multiple Linear Regression and Variable Selection 2, hereafter it will be referred to as Model 2).
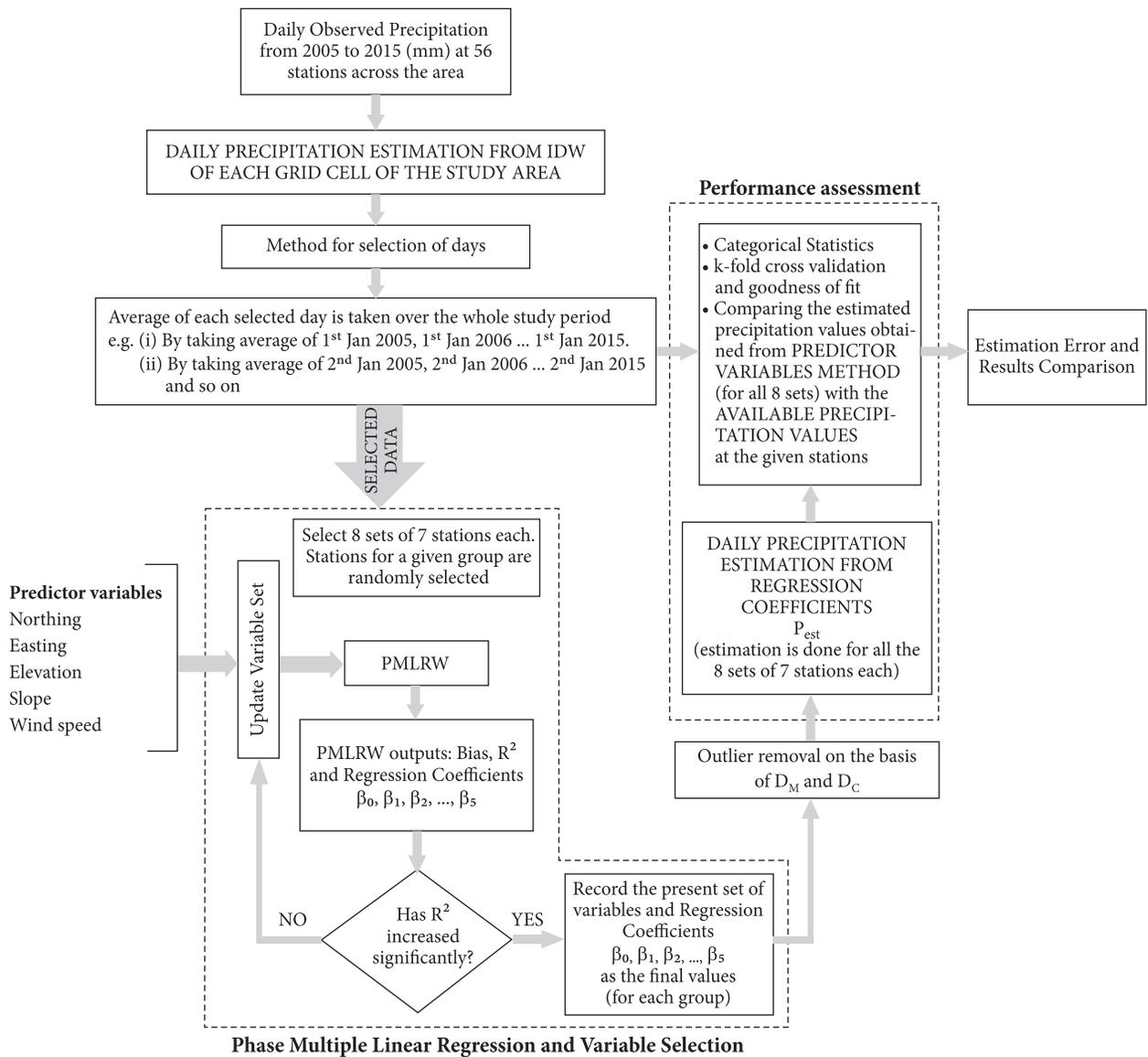
Figure 3. The general proposed methodology which was applied for both the models

1.2.4. The predictor variables selection criteria

Since we have the average of each day throughout the study period, the data is separated month wise because of monthly dynamic nature of precipitation. Now for a particular month, all the possible combination of predicted variables were assessed through MLR. For a particular month a variable is added and then MLR is applied. If R^2 is increased 5% or more than 5%, then this variable is selected for that particular month. Now we add another variable and MLR is conducted, if again R^2 will increase by more than 5%, this variable added for this month else it is not added. In the same way keep on adding variables and finally a combination of variables for a particular month is formed such that on addition of each of these variables results in more than 5% increase in R^2. All other variables for this month are rejected. The same process will applied on all months.

Apart from the month wise operation, the whole process is applied for season wise as well as year wise to know the behaviour of the predictor variables in both the models in these situations as well. The choice of 5% increase in R^2 is taken because increment affects the Normal P-P plot to a considerable amount and the error between observed and estimated precipitation (Figure S4).

1.2.5. Performance assessment

Performance assessment of Model 1 and Model 2 was done by Categorical statistics, Goodness-of-fit and *k*-fold cross validation methods. The categorical statistics were used to assess the phase estimation which gives binary predictions of occurrence or non-occurrence of precipitation as shown in Table 1.

In this study, we used five categorical statistics as those employed by Castro et al. (2014) including frequency bias as shown in Table 2.

Table 1. Confusion matrix which is used to define categorical measures for the occurrence/non-occurrence of estimated precipitation for Model 1 and Model 2

	Number (No.) of wet Phase in Estimation	No. of dry Phase in Estimation
No. of wet Phase in Observation	a	d
No. of dry Phase in Observation	b	c

Table 2. Categorical statistics to asses phase estimation of precipitation. The value 1 is for perfect prediction

Statistics	Equations
Proportion Correctly Predicted (PCP)	$PCP = \dfrac{a+b}{a+b+c+d}$
Probability of Detection (POD)	$POD = \dfrac{a}{a+c}$
Critical Success Index-Wet (CSI$_{wet}$)	$CSI_{wet} = \dfrac{a}{a+b+c}$
Critical Success Index-Dry (CSI$_{dry}$)	$CSI_{dry} = \dfrac{d}{b+c+d}$
Frequency Bias	$Bias = \dfrac{a+b}{a+c}$

Goodness-of-fit was assessed by two steps. The first is the removal of outlier (an observation that lies at an abnormal distance from other distributed values) on the basis of Mahalanobis (D_M) and Cook's distance (D_C). Mahalanobis distance (Mahalanobis, 1936) is used to measure how far a point measurement is from the centroid of a reference distribution (Hamill, Giordano, Ward, Giles, & Holben, 2016) whereas Cook's distance (Cook, 1977) is used to estimate the influence of removal of a data point on distribution during least-squares regression analysis (Pinho, Nobre, & Singer, 2015). Second is Bias which is simply the mean of the residuals, indicating whether the model tends to under or over-estimate the measured data, with an ideal value of zero (Bennett et al., 2013). These methods were conducted for each rain gauge, and are given by:

$$D_M = \left[\left(x - m \right)^T S^{-1} \left(x - m \right) \right]^{1/2}; \qquad (7)$$

$$D_C = \frac{E_i^2}{s^2 p} \left[\frac{h_i}{\left(1 - h_i\right)^2} \right]; \qquad (8)$$

$$Bias = \frac{1}{n} \sum_{j=1}^{n} \left[P_{est.j} - P_{obs.j} \right], \qquad (9)$$

where x is the distance of observation; m is the cluster of mean values; S is the covariance matrix; E_i is standardized residual; h_i is leverage; s is mean squared error; i is an integer; p is the number of coefficient of regression model; n

is number of rain gauges; $P_{est.j}$ and $P_{obs.j}$ are estimated and observed precipitation respectively for each day j.

k-fold cross-validation is a widely used method to assess the performance of model. Data are divided into k sets and one set is used for training purpose (Bennett et al., 2013). The remaining k–1 sets are used for model development. The MLR is applied on k–1 set and then precipitation is estimated for the training set with the help of the coefficients obtained from k–1 sets. The process is then repeated k times and average bias across all the k trials is computed.

2. Results and discussion

The two fold daily spatial estimation of precipitation was applied on average daily basis of the study period and Model 1 with Model 2 were compared to determine the performance of these models in tropical climate. In cross-validation process, $k = 8$ sets were selected in which each set has 7 randomly chosen rain gauge stations without repetition, leaving the other 7 sets (49 stations) for interpretation.

2.1. Phase estimation

In precipitation estimation, the range of R_{inf} for Eq. (2) was taken from 10 km to 90 km. Incremental steps of 10 km each were taken until at least 3 rainfall stations were located within the radius. In the second method for phase estimation, a range of 1.5 and 5 were taken with incremental steps of 0.5 on two first two stations to calculate b value in Eq. (5). The categorical statistics were then applied on these two stations to ascertain the changes in b value. No convincing difference was found in the categorical statistics of these values, therefore $b = 2$ was chosen for the rest of the study.

After the phase estimation for each grid cell, categorical statistics were applied to assess the overall performance of occurrence/non-occurrence of estimated precipitation. Figure 4 shows the performance of the categorical statistics by plotting their mean monthly values for both the models. PCP for Model 1 was ranged from 0.68 to 0.78 while for Model 2 from 0.62 to 0.81 but the average values are 0.73 and 0.71 respectively which is quite similar for both the models (Figure 4a). Highest values are from November and December for both the models reflecting better performance in high precipitation. Slightly lower values of PCP for Model 2 during February to March are caused by wet phase in estimation and dry phase in observation (false alarms) that reflect incapability of Model 2 to recognise local storms which is just after the monsoon season.

POD, which is sensitive to hits, produces fairly constant values which ranged from 0.90 to 0.97 and 0.99 to 1 for Model 1 and Model 2 respectively (Figure 4b) showing perfect model for identifying wet phase in Model 2. However, Model 2 failed to detect in CSI$_{dry}$ as evident from the miniscule variation throughout the year ranging from 0.01
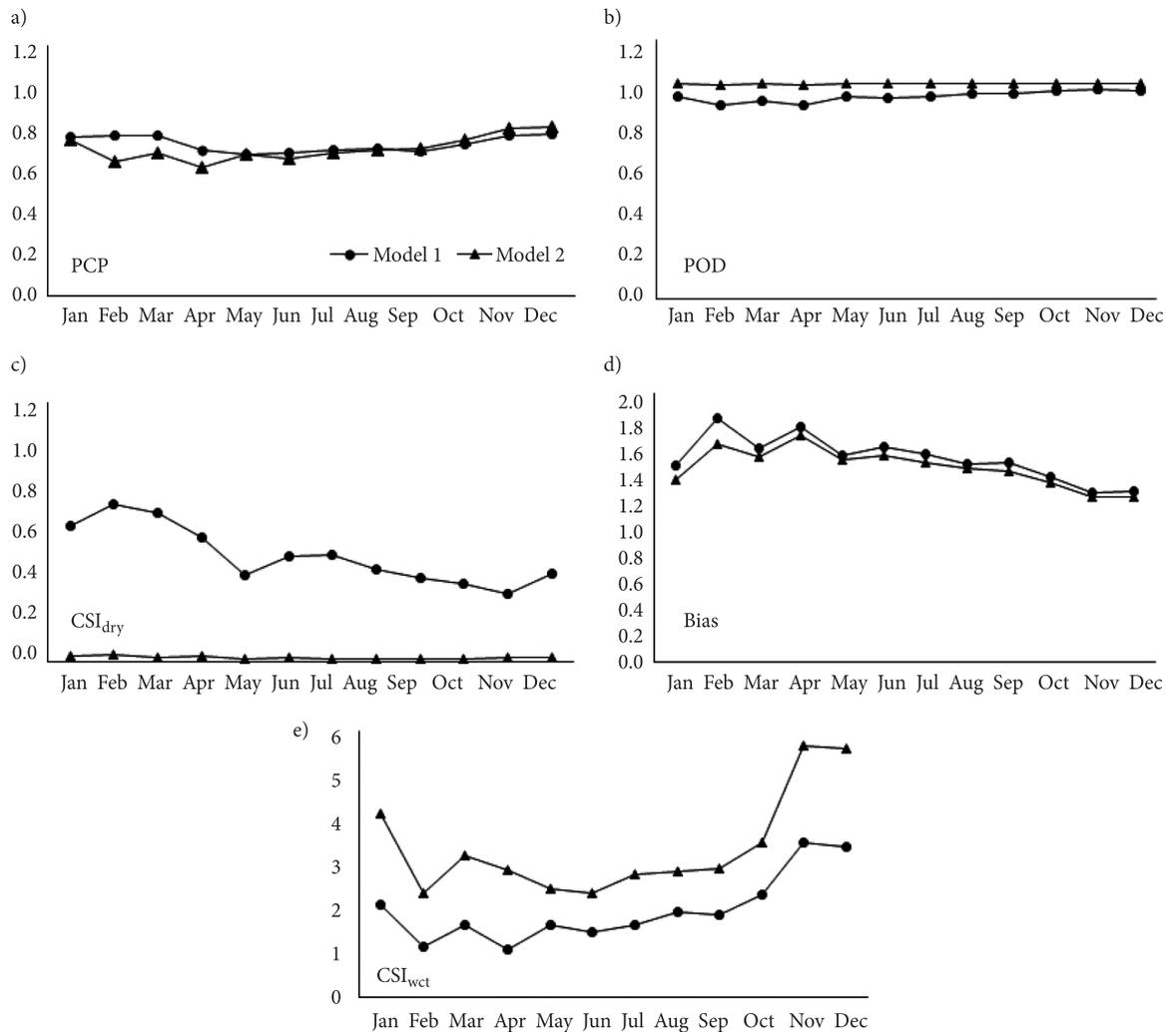
Figure 4. Monthly variation of categorical statistics: a) PCP, b) POD, c) CSI_{dry}, d) Bias and e) CSI_{wet}

and 0.03 (Figure 4c). Only in this particular case, Model 1 performed better ranging from 0.3 to 0.71 with lowest value in November. Such reduction is associated with the significant number of false alarms for the entire period previously interpreted (Castro et al., 2014). Although, both the models performed the poorest in CSI_{wet} with average values 1.97 and 3.39 for Model 1 and Model 2 respectively (Figure 4e). Furthermore, the overall performance of Model 2 is better than Model 1 (Figure 4d) but they showed similar behaviour in November and December with average values of 1.29 and 1.25 respectively.

2.2. Interpolation method and its overall performance

The predictor variables are selected in Eq. (6) with the assumption of linear dependency of precipitation on these variables. Figure 5 and Figure 6 shows box-and-whisker plots for the overall behaviour of regression coefficients and R^2 respectively obtained from both the models. The data to build the box plots for each rain gauge in validation partitioning method are associated with average set of regression coefficients and R^2 for average daily linear regression of month wise interpolation. The season wise and year wise interpolation have not include in this because of missing some variables as per the variable selection criteria. The slope variable was drop for this study because it was not contributed as per the selection criteria in all interpolation cases.

The length of each box indicates the interquartile range, the middle line is the median and the whiskers are 25th and 75th percentiles. The resulting regression coefficients for Model 1 and Model 2 are similar except for easting and elevation which is showing high values of both negative and positive errors in Model 2 and the more variation in 25th and 75th percentile of elevation for both the models respectively (Figure 5) whereas, R^2 values are higher in Model 1 (Figure 6). Average minimum value of R^2 for Model 1 is 0.51 and maximum is 0.76 with median of 0.67. Compared to this, Model 2 has quite low values of minimum i.e. 0.36 to maximum of 0.66 with median of 0.44. High variability of coefficients in Model 2 shows high range of estimated values. Due to high variation in coefficients, Model 2 has low R^2 values. Therefore, the results clearly indicates that the overall performance of Model 1 is better than Model 2.
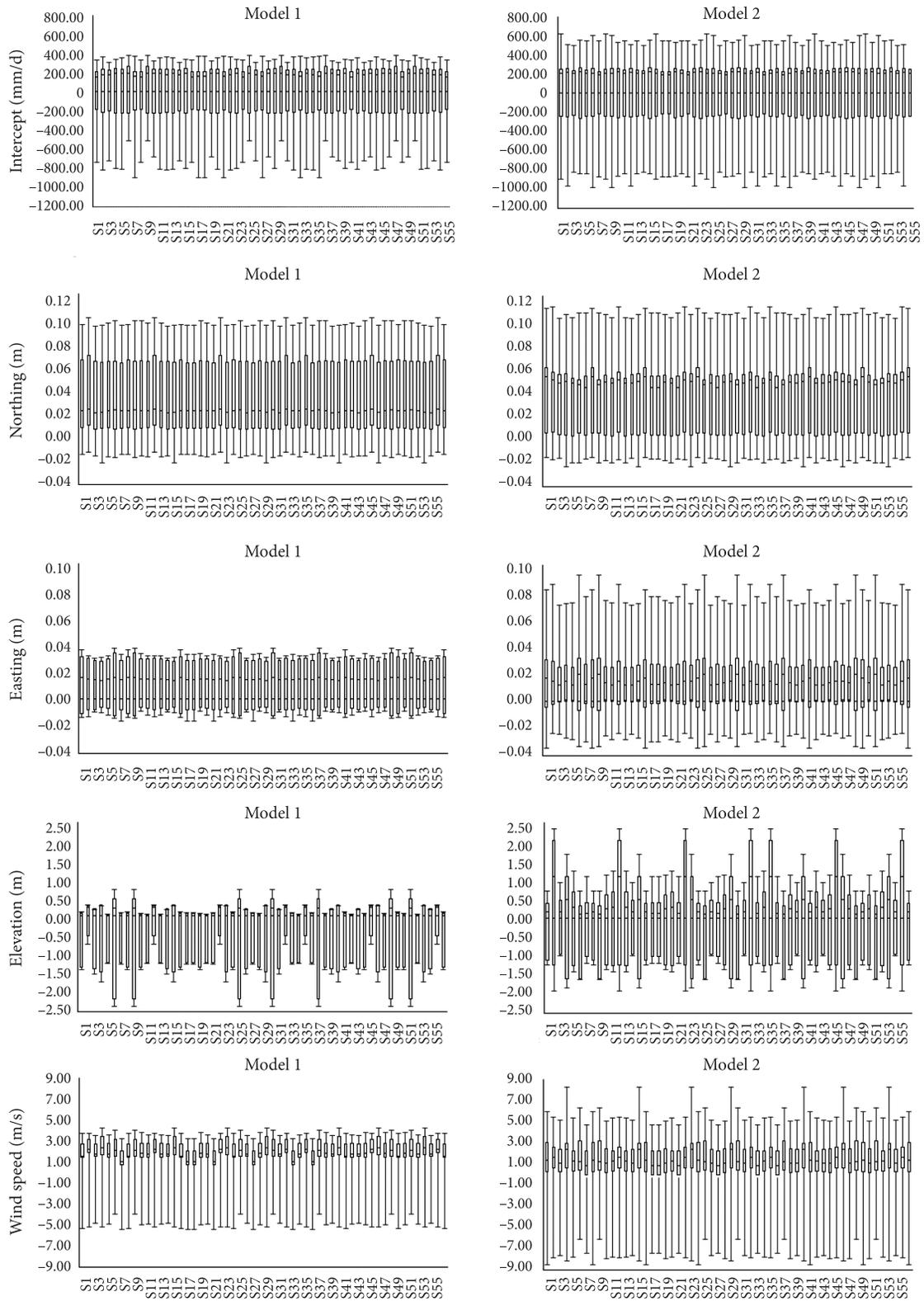
Figure 5. Box-and-whisker plot for regression coefficients for overall monthly performance obtained after cross validation. X-axis is showing the station numbers
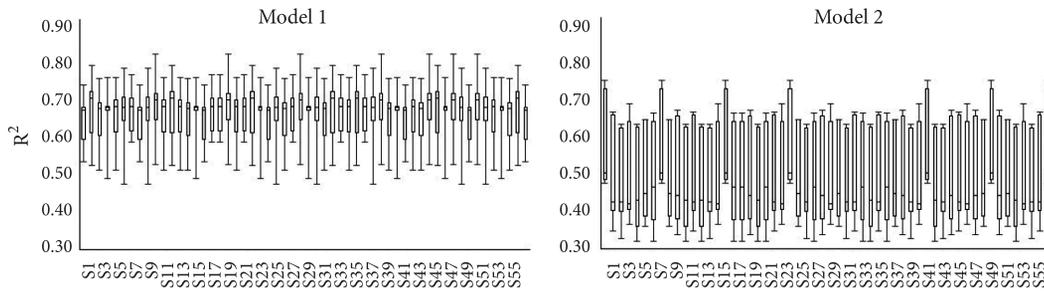
Figure 6. Box-and-whisker plots for overall monthly performance of R2 obtained from k-fold cross

2.3. Month wise performance of interpolation methods

In variable selection, Northing (N), Easting (Ea) and wind speed (W) dominated for both the models in this case. Table 3 shows the dominancy of Ea and W in Model 2 and N in Model 1 while elevation (E) shows a balanced performance in both the models. The results also conclude that N perform lonely in high precipitation such as January, November and December in Model 1 and January in Model 2. The role of W is found to be very important as it increases the R^2 values for both the models. September shown only 7% increment while others are greater than 10% with maximum 54% for February.

The goodness of fit assessment shows the Model 1 performed better than as Model 2 in October and vice versa in January. The bias, in this case, of both the models also show the similar results and there are only slight overall differences in average values (i.e. 0.016 for Model 1 and 0.010 for Model 2). The bias was obtained from *k*-fold partitioning cross validation method. The overall performance of Model 1 and Model 2 are similar in terms of both goodness-of-fit and bias values while Model

1 performed better than Model 2 in terms of R^2 values and error between observed and estimated values. The Normal probability-probability (P-P) plots based on the standardized residuals for each month of both the models are shown in Figure S2.

2.4. Season wise performance of interpolation methods

The maximum rainfall of the study area is mainly from November to January but for the entire country of Malaysia, high rainfall can be seen from October to January and in this case February is also included in the season because of its high precipitation in some years. Hence, October to February is referred to as wet season while the other season from March to September is referred to as dry season. Different combination of months in seasonal data were used. The procedure of data separation was same as month wise interpolation. The interpolation was performed on wet season because in dry season, there is poor effect of predictor variables in the increment of R^2 in all combinations of months (Table S1).

Table 3. Performance of Model 1 and Model 2 for month wise interpolation. P_{mean} values, bias, R^2 values range in terms of variable performance and their contribution in R^2 increment. Percent rise in R^2 is shown by last variable in each case

Month	Model 1					Model 2				
	P_{mean} (mm/d)	R^2	Bias (mm/d)	Variables	% rise in R^2	P_{mean} (mm/d)	R^2	Bias (mm/d)	Variables	% rise in R^2
January	15.46	0.68	0.000	N	–	13.22	0.58	0.000	N	–
February	9.42	0.60	−0.010	N.EaW	54	7.45	0.52	−0.047	N.Ea.W	29
March	10.70	0.61	0.021	Ea.W	43	8.81	0.48	0.015	Ea.W	28
April	8.29	0.55	0.045	N.E.W	45	6.84	0.52	−0.109	N.E.W	48
May	9.93	0.52	0.012	Ea.E.W	16	8.50	0.50	0.010	N.Ea.E	10
June	9.62	0.48	0.035	Ea.W	16	8.16	0.49	0.036	Ea.W	19
July	8.89	0.48	0.024	N.E.W	28	7.59	0.42	0.008	Ea.E.W	13
August	10.42	0.48	0.003	N.Ea	11	9.15	0.48	0.039	Ea.W	23
September	10.31	0.35	0.022	N.W	7	9.00	0.48	0.025	Ea.W	15
October	11.23	0.35	0.034	Ea.W	13	10.16	0.44	0.025	N.E.W	32
November	18.50	0.70	0.002	N	–	17.49	0.68	0.081	N.Ea.W	7
December	23.90	0.83	0.003	N	–	22.02	0.77	0.034	N.W	7

The combinations of wet season months are showing in Table 4. In variable selection, N is found to dominate in both the models but its effect is more in Model 1 (Table 4). N performed in 7 out of 10 cases alone in Model 1 while 3 out of 10 in Model 2. In this case, W also performs well in R^2 increment which ranges from 5% to 22% in Model 1 while 5% to 18% in Model 2. Highest R^2 value (0.79) was observed in "December to February" for Model 1 whereas $R^2 = 0.71$ in "December and January" for Model 2. Therefore, season wise performance of Model 1 is also found better in this situation as compared to Model 2.

In goodness of fit assessment, Model 1 is showing disturbance in normal plot for "December to February", "December & January" and "January & February" as compared to Model 2. The best performance of Model 1 and Model 2 are shown by "October & November" which have minimum R^2 in both the models. The Model 1 has not

shown much variation which is showing the good performance of Model 1 in terms of bias. Again, it can be concluded that the better performance for season wise combinations of Model 1 than Model 2 in terms of bias. The Normal P-P plot for all season wise combinations of both the models is shown in Figure S3.

2.5. Year wise performance of interpolation methods

Variables such as N, Ea and E in this case are found to have similar contribution but still W dominates in R^2 increment in both the models (Table 5). The range of percent rise by W in R^2 is 5% to 14% for Model 1 whereas 7% to 23% for Model 2. In this case too, N performed alone in the years 2010 and 2011 for both the models but worst performance by all variables in the year of 2014 and 2015

Table 4. Performance of Model 1 and Model 2 for season wise interpolation. P_{mean} values, bias, R^2 values range in terms of variable performance and their contribution in R^2 increment. Percent rise in R^2 is shown by last variable in each case

Months combinations	Model 1					Model 2				
	P_{mean} (mm/d)	R^2	Bias (mm/d)	Variables	% rise in R^2	P_{mean} (mm/d)	R^2	Bias (mm/d)	Variables	% rise in R^2
October-February	14.11	0.71	0.002	N	–	15.80	0.65	−0.005	N.W	5
October-January	15.74	0.73	0.001	N	–	17.25	0.64	0.001	N	–
November-February	15.08	0.74	0.002	N	–	16.89	0.69	0.021	N.W	6
October-December	16.57	0.73	0.002	N	–	17.94	0.69	0.012	N.Ea	–
November-January	17.57	0.78	0.002	N	–	19.19	0.69	−0.056	N	–
December-February	14.33	0.79	0.013	N.W	10	16.36	0.68	0.006	N.W	16
October & November	13.85	0.66	0.044	N.W	5	15.01	0.65	0.062	N.Ea.W	7
November & December	19.72	0.78	0.003	N	–	21.20	0.71	0.000	N	–
December & January	17.70	0.78	0.001	N	–	19.54	0.71	0.029	N.W	6
January & February	10.48	0.70	0.007	N.W	22	12.55	0.58	−0.008	N.W	18

Table 5. Performance of Model 1 and Model 2 for year wise interpolation. P_{mean} values, bias, R^2 values range in terms of variable performance and their contribution in R^2 increment. Percent rise in R^2 is shown by last variable in each case

Year	Model 1					Model 2				
	P_{mean} (mm/d)	R^2	Bias (mm/d)	Variables	% rise in R^2	P_{mean} (mm/d)	R^2	Bias (mm/d)	Variables	% rise in R^2
2005	10.14	0.50	0.043	Ea.W	14	11.50	0.47	0.048	Ea.W	22
2006	10.80	0.47	0.058	N.W	5	12.08	0.59	0.049	N.W	7
2007	11.90	0.37	0.004	N.W	8	13.19	0.41	−0.216	N.W	13
2008	10.02	0.54	0.035	N.Ea.W	6	11.72	0.53	0.025	N.Ea	12
2009	12.11	0.67	−0.015	N.Ea	7	13.66	0.65	0.008	N.Ea	8
2010	9.47	0.38	0.000	N	–	11.18	0.35	0.006	N	–
2011	11.66	0.47	−0.004	N	–	12.97	0.48	−0.002	N	–
2012	9.67	0.41	−0.004	Ea.E.W	14	11.10	0.58	0.044	Ea.W	23
2013	10.55	0.28	−0.031	Ea.E	6	12.28	0.41	0.008	Ea	–
2014	11.40	0.03	0.003	E	–	13.48	0.02	0.019	E	–
2015	7.90	0.06	−0.010	E	–	9.76	0.07	−0.012	E	–

is observed. Highest R^2 in the worst condition is shown by E as shown in Table 5. The best R^2 is shown in 2009 for both the models. In terms of R^2, Model 1 also performs better than Model 2 in this case.

In this case, both the models shows similar results for each year. Similarly, bias provided approximately the same results instead of the expected negative values which are much higher in Model 2 (−0.22) as compared to Model 1 (−0.03). Therefore, Model 1 also perform better than Model 2 in terms of bias.

As the interpolation was applied on average daily precipitation, it can be noted that the importance of group of predictor variables decreases from month wise to season wise and year wise. It also indicate that in high precipitation months especially in the northeast monsoon N dominates in precipitation estimation.

2.6. Spatial interpolation of daily precipitation and validation

To know the performance of spatial interpolation on daily precipitation estimation, 17th and 22nd December 2014 precipitation were selected. These are the two highest events in 2014 flood in the area. To generate daily precipitation maps for these events, apart from both the models, IDW was also used to interpolate observed precipitation values. These interpolations were done to understand the spatial behaviour of variables and their role

in spatial estimation of daily precipitation through Model 1 and Model 2. Although, the method of finding wet days for both the models are different, however, on these dates both the models show same amount of precipitation. As it is shown in month wise interpolation, N dominates in December in Model 1 while N and W dominates in Model 2 in the estimation of average daily precipitation. Hence, same variables were included for daily precipitation interpolation because the selected days are from the month of December. The maps for daily interpolations and their results are shown in Figure 7 and Table 6 respectively.

By comparing the results, the average spatial estimation of precipitation shows closer estimated values by both

Table 6. The minimum, maximum and average values from IDW, Model 1 and Model 2

	IDW (mm/d)		Model 1 (mm/d)		Model 2 (mm/d)	
	17th December	22nd December	17th December	22nd December	17th December	22nd December
Minimum	19.9	1.2	12.9	1	10.7	1
Maximum	1137	478.7	332.5	285.5	333	258.4
Average	146.3	164.9	151.1	155.6	161.9	158.5
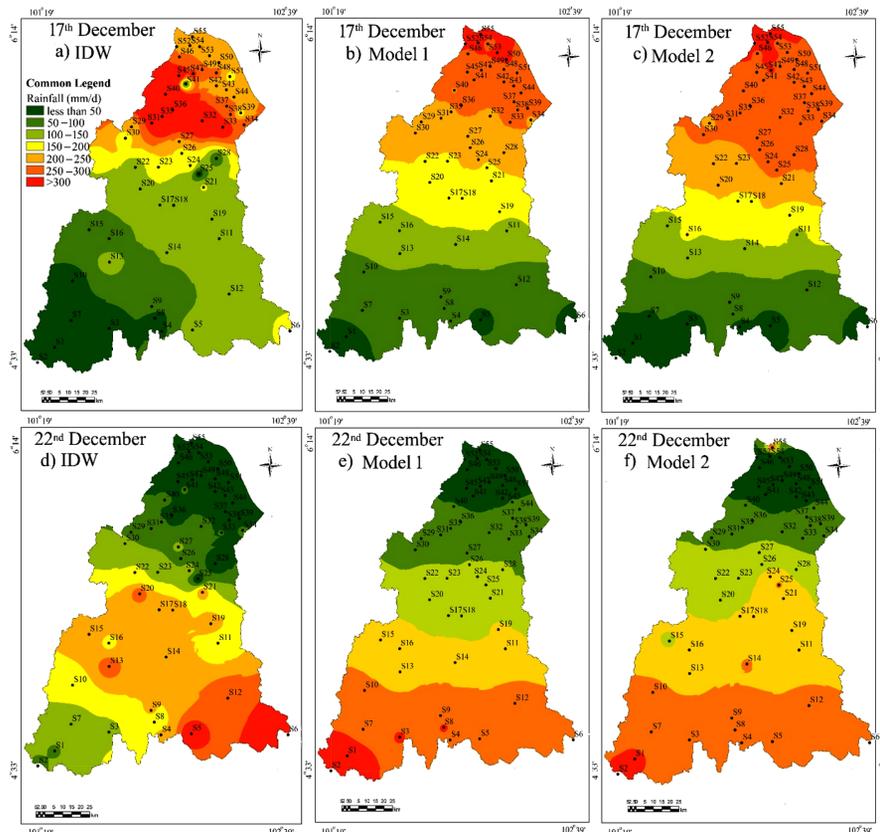
Figure 7. a), b), c) and d), e) and f) are spatial distribution of precipitation estimated by IDW, Model 1 and Model 2 respectively for 17th and 22nd
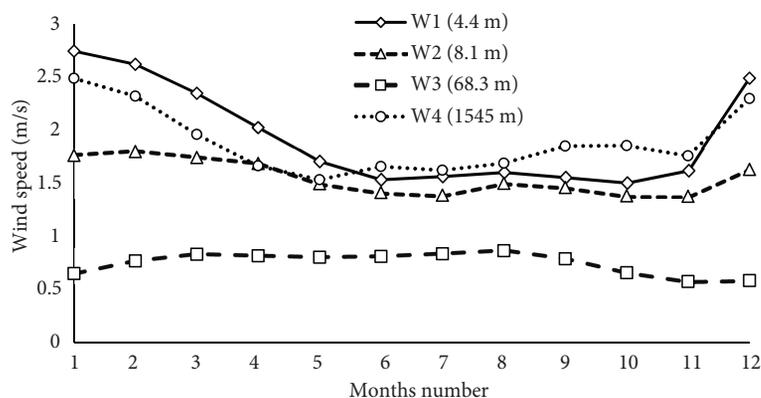
Figure 8. Average daily wind speed pattern for the whole year in the area

the models with underestimation on 17th and 22nd December. The inclusion of wind speed has decreased the error between observed and estimated average values on 22nd December.

Daily spatial estimation of precipitation by IDW method on both days produced some random precipitation distribution in the area. Figure 7(a) is showing low precipitation at upstream and midstream which comes under high to moderate relief areas while Figure 7(d) is showing low precipitation at downstream. As compared to IDW, Model 1 and Model 2 have produced zones of constant precipitation covering different relief. These zones are reflecting northing which proves the dominating behaviour of the predictor variable during high precipitation. It can be seen from month wise and season wise interpolation performance by Model 1 and Model 2 in the previous sections. Figure 7(c) is showing effect of wind speed in which high precipitation values were slightly shifted from the north to the south direction. The wind speed pattern values on 17th and 22nd December were high at downstream and low at upstream. The wind speed on 17th December (maximum 4.1 m/s) was low as compared to 22th December (maximum 7.0 m/s). Therefore, the precipitation shift from north to south is showing the effect of wind speed and also supporting the climatic condition of the study area.

In Figure 7(d) and 7(e), high precipitation was observed at S6 (at 976 m elevation) while in the estimation, high precipitation was at S1 and S2 (at 624 m and 1707 m elevation respectively). This was the worst estimation on 22nd December. However, IDW is clearly not showing any reflection of predictor variables. Figure 7(f) is showing that the high precipitation values were slightly shifted from south to north direction because of high wind speed at downstream and low wind speed at upstream as compared to 17th December. The shifting pattern indicates that along with northing, wind speed is playing an important role in spatial distribution of precipitation in the area. Therefore, it can be concluded that the performance of Model 1 and Model 2 in spatial distribution of precipitation is better as compared to IDW.

The wind speed behaviour in Figure 8 is showing higher values closer to shore line station (W1) and it is

decreasing towards midstream and then increases at high relief station (W4).

It is indicating the effect of distance from shore line on wind characteristics. In this situation, wind direction along with shore line can play important role in the understanding of the behaviour of wind characteristic in precipitation occurrence. Although, slopes did not show good correlation in both the models but they might develop better correlation with wind direction. Therefore, different results of spatial distribution and magnitude of error strengthens the conclusion that interpolation methods should be assessed for slope and wind direction in mountainous regions where precipitation are being poorly measured.

Conclusions

The simple threshold method for phase estimation is found better in terms of simple threshold assumption in which greater than zero precipitation is taken as wet days because of the typical nature of consistent precipitation in tropical climate throughout the year. The phase estimation of precipitation improves the performance of both the models but Model 2 slightly performed better than Model 1 in terms of categorical statistics. Relationship between precipitation and different predictor variables was shown the dominancy of northing in approximately all cases. Wind speed has shown very important role in the increment of R^2 in most of the cases while easting behaviour was found better as compared to elevation. Although the area has only four wind speed stations, this might have resulted in lower accuracy. But since the methodology is general, it can be applied to any region. This shows that northing and wind speed plays very important role in the interpolation of precipitation. The overall interpolation of precipitation for month wise, season wise and year wise performance of Model 1 was better than Model 2 in terms of R^2 and bias but both the models produce mixed results in goodness of fit assessment. The best performance was observed for both the models in season wise and in those months where precipitation was high (i.e. December to February and November & December). Maximum R^2

value was observed for the month of December (i.e. 0.83 and 0.77 for Model 1 and Model 2 respectively) and for the combined months of December to February (0.79) for Model 1. However, both the models overestimated precipitation but Model 2 overestimation was higher as compared to Model 1.

Spatial distribution of observed and estimated precipitation patterns of the two flooding days of December 2014 in the area conclude that both the models produced zones which are validating the dominancy of northing in special distribution of precipitation in the area. A slight northing effect of wind speed in shifting spatial distribution of precipitation was also observed in Model 2. Special patterns from IDW shows lack of predictor variables effect on special distribution of precipitation. The overall results show that the northing, easting, elevation and wind characteristics play very important role in special precipitation distribution which is a key input for hydrological modeling in flood risk and soil loss estimation. Whereas, slope neither shows good correlation nor does it affects the precipitation estimation. However, the effect of slope on the wind direction may be an important variable in precipitation estimation as well as revealed the behaviour of elevation and wind speed in the area. The results from this study highlight the importance of future researches that should include wind direction, relative humidity and distance from shoreline at hourly precipitation estimation, as they affect the rise of air masses in this climate. The MLR should be compared with other methods such as ARMAX (Autoregressive–moving-average model with exogenous inputs) and Neural Networks (ANN). Furthermore, on the basis of the study, it is suggested that more wind characteristics station should be installed in the area to enhance the accuracy in calculating wind characteristics variables in each grid cell of the area.

Acknowledgements

References

Allamano, P., Claps, P., Laio, F., & Thea, C. (2009). A data-based assessment of the dependence of short-duration precipitation on elevation. *Physics and Chemistry of the Earth*, *34*(10), 635-641. https://doi.org/10.1016/j.pce.2009.01.001

Anees, M. T., Abdullah, K., Nawawi, M. N. M., Rahman, N. N. N. A., Piah, A. R. M., Omar, F. M., Syakir, M. I.,

Zakaria, N. A., & Abdul Kadir, M. O. (2017). Development of daily rainfall erosivity model for Kelantan state, Peninsular Malaysia. *Hydrology Research*, *49*(5), 1434-1451. https://doi.org/10.2166/nh.2017.020

Bennett, N. D., Croke, B. F., Guariso, G., Guillaume, J. H., Hamilton, S. H., Jakeman, A. J., Marsili-Libelli, S., Newham, L. T., Norton, J. P., Perrin, C., & Pierce, S. A. (2013). Characterising performance of environmental models. *Environmental Modelling & Software*, *40*, 1-20. https://doi.org/10.1016/j.envsoft.2012.09.011

Buytaert, W., Celleri, R., Willems, P., De Bievre, B., & Wyseure, G. (2006). Spatial and temporal rainfall variability in mountainous areas: A case study from the south Ecuadorian Andes. *Journal of Hydrology*, *329*(3), 413-421. https://doi.org/10.1016/j.jhydrol.2006.02.031

Carpenter, T. M., & Georgakakos, K. P. (2004). Impacts of parametric and radar rainfall uncertainty on the ensemble streamflow simulations of a distributed hydrologic model. *Journal of Hydrology*, *298*(1), 202-221. https://doi.org/10.1016/j.jhydrol.2004.03.036

Castro, L. M., Gironás, J., & Fernández, B. (2014). Spatial estimation of daily precipitation in regions with complex relief and scarce data using terrain orientation. *Journal of Hydrology*, *517*, 481-492. https://doi.org/10.1016/j.jhydrol.2014.05.064

Cook, R. D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*(1), 15-18. https://doi.org/10.1080/00401706.1977.10489493

Daly, C., Neilson, R. P., & Phillips, D. L. (1994). A statistical-topographic model for mapping climatological precipitation over mountainous terrain. *Journal of applied meteorology*, *33*(2), 140-158. https://doi.org/10.1175/1520-0450(1994)033<0140:ASTMFM>2.0.CO;2

Diodato, N. (2005). The influence of topographic co-variables on the spatial variability of precipitation over small regions of complex terrain. *International Journal of Climatology*, *25*(3), 351-363. https://doi.org/10.1002/joc.1131

Drogue, G., Humbert, J., Deraisme, J., Mahr, N., & Freslon, N. (2002). A statistical–topographic model using an omnidirectional parameterization of the relief for mapping orographic rainfall. *International Journal of Climatology*, 22(5), 599-613. https://doi.org/10.1002/joc.671

Gonga-Saholiariliva, N., Neppel, L., Chevallier, P., Delclaux, F., & Savéan, M. (2016). Geostatistical estimation of daily monsoon precipitation at fine spatial scale: Koshi river basin. *Journal of Hydrologic Engineering*, *21*(9), 1-15. https://doi.org/10.1061/(ASCE)HE.1943-5584.0001388

Goovaerts, P. (2000). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, *228*(1), 113-129. https://doi.org/10.1016/S0022-1694(00)00144-X

Guan, H., Wilson, J. L., & Makhnin, O. (2005). Geostatistical mapping of mountain precipitation incorporating autosearched effects of terrain and climatic characteristics. *Journal of Hydrometeorology*, *6*(6), 1018-1031. https://doi.org/10.1175/JHM448.1

Hamill, P., Giordano, M., Ward, C., Giles, D., & Holben, B. (2016). An AERONET-based aerosol classification using the Mahalanobis distance. *Atmospheric Environment*, *140*, 213-233. https://doi.org/10.1016/j.atmosenv.2016.06.002

Hewitson, B. C., & Crane, R. G. (2005). Gridded area-averaged daily precipitation via conditional interpolation. *Journal of Climate*, *18*(1), 41-57. https://doi.org/10.1175/JCLI3246.1

Hwang, Y., Clark, M., Rajagopalan, B., & Leavesley, G. (2012).

Spatial interpolation schemes of daily precipitation for hydrologic modeling. *Stochastic Environmental Research and Risk Assessment*, 26(2), 295-320. https://doi.org/10.1007/s00477-011-0509-1

Johansson, B., & Chen, D. (2003). The influence of wind and topography on precipitation distribution in Sweden: Statistical analysis and modelling. *International Journal of Climatology*, 23(12), 1523-1535. https://doi.org/10.1002/joc.951

Johnson, F., Hutchinson, M. F., The, C., Beesley, C., & Green, J. (2016). Topographic relationships for design rainfalls over Australia. *Journal of Hydrology*, 533, 439-451. https://doi.org/10.1016/j.jhydrol.2015.12.035

Kurtzman, D., Navon, S., & Morin, E. (2009). Improving interpolation of daily precipitation for hydrologic modelling: spatial patterns of preferred interpolators. *Hydrological Processes*, 23(23), 3281-3291. https://doi.org/10.1002/hyp.7442

Ly, S., Charles, C., & Degre, A. (2011). Geostatistical interpolation of daily rainfall at catchment scale: the use of several variogram models in the Ourthe and Ambleve catchments, Belgium. *Hydrology and Earth System Sciences*, 15(7), 2259-2274. https://doi.org/10.5194/hess-15-2259-2011

Mahalanobis, P. (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2, 49-55.

Mair, A., & Fares, A. (2010). Comparison of rainfall interpolation methods in a mountainous region of a tropical island. *Journal of Hydrologic Engineering*, 16(4), 371-383. https://doi.org/10.1061/(ASCE)HE.1943-5584.0000330

Masseran, N., & Razali, A. M. (2016). Modeling the wind direction behaviors during the monsoon seasons in Peninsular Malaysia. *Renewable and Sustainable Energy Reviews*, 56, 1419-1430. https://doi.org/10.1016/j.rser.2015.11.040

McCuen, R. H. (1989). *Hydrologic analysis and design*. N.J.: Prentice-Hall Englewood Cliffs.

Mikoš, M., Jošt, D., & Petkovšek, G. (2006). Rainfall and runoff erosivity in the alpine climate of north Slovenia: a comparison of different estimation methods. *Hydrological Sciences Journal*, 51(1), 115-126. https://doi.org/10.1623/hysj.51.1.115

Pinho, L. G. B., Nobre, J. S., & Singer, J. M. (2015). Cook's distance for generalized linear mixed models. *Computational Statistics & Data Analysis*, 82, 126-136. https://doi.org/10.1016/j.csda.2014.08.008

Qing, Y., Zhu-Guo, M. A., & Liang, C. (2011). A preliminary analysis of the relationship between precipitation variation trends and altitude in China. *Atmospheric and Oceanic Science Letters*, 4(1), 41-46. https://doi.org/10.1080/16742834.2011.11446899

Rajagopalan, B., & Lall, U. (1998). Locally weighted polynomial estimation of spatial precipitation. *Journal of Geographic Information and Decision Analysis*, 2(2), 44-51. Retrieved from https://www.researchgate.net/profile/Upmanu_Lall/publication/222797769_Locally_weighted_polynomial_estimation_of_spatial_precipitation/links/0fcfd50943f2a6988c000000.pdf

Seo, D. J. (1998). Real-time estimation of rainfall fields using rain gage data under fractional coverage conditions. *Journal of Hydrology*, 208(1), 25-36. https://doi.org/10.1016/S0022-1694(98)00140-1

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 23rd ACM national conference, ACM* (pp. 517-524). New York, NY, USA: ACM. https://doi.org/10.1145/800186.810616

Tabios, G. Q., & Salas, J. D. (1985). A comparative analysis of techniques for spatial interpolation of precipitation. *JAWRA*, 21(3), 365-380. https://doi.org/10.1111/j.1752-1688.1985.tb00147.x

Tao, T. (2009). Uncertainty analysis of interpolation methods in rainfall spatial distribution–a case of small catchment in Lyon. *Journal of Environmental Protection*, 1(01), 50-58.

Thiessen, A. H. (1911). Precipitation averages for large areas. *Monthly Weather Review*, 39(7), 1082-1089. https://doi.org/10.1175/1520-0493(1911)39<1082b:PAFLA>2.0.CO;2

Xie, P., Chen, M., Yang, S., Yatagai, A., Hayasaka, T., Fukushima, Y., & Liu, C. (2007). A gauge-based analysis of daily precipitation over East Asia. *Journal of Hydrometeorology*, 8(3), 607-626. https://doi.org/10.1175/JHM583.1

Appendix

Table S1. Performance of Model 1 and Model 2 in dry season

Months combinations	P_{mean} (mm/d)	R^2
March-September	9.8	0.26
March-August	9.7	0.27
April-September	9.6	0.12
March-July	8.0	0.39
April-August	8.1	0.22
May-September	8.5	0.25
March-June	8.1	0.43
April-July	7.8	0.27
May-August	8.3	0.24
June-September	8.5	0.26
March-May	8.1	0.42
April-June	7.8	0.28
May-July	8.1	0.30
June-August	8.3	0.24
July-September	8.6	0.19
March-April	7.9	0.50
April-May	7.7	0.25
May-June	8.3	0.34
June-July	7.9	0.33
July-August	8.4	0.15
August-September	9.1	0.18
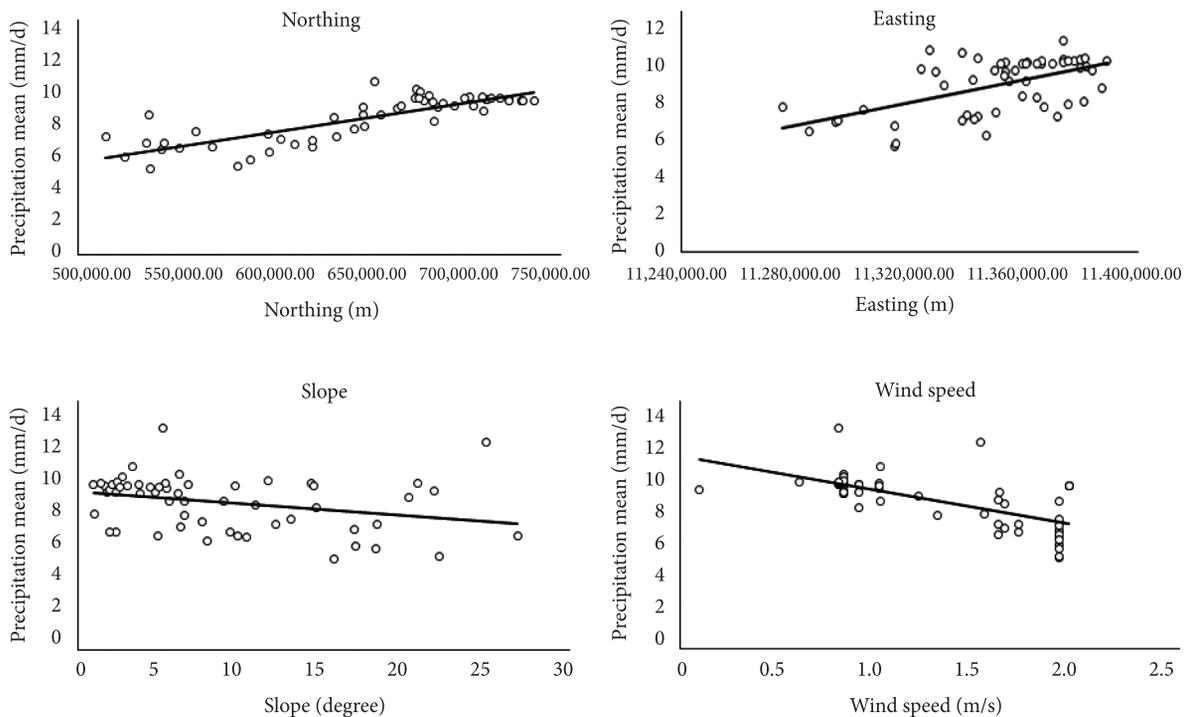
Figure S1. The relationship between precipitation mean with northing, easting, slope and wind speed
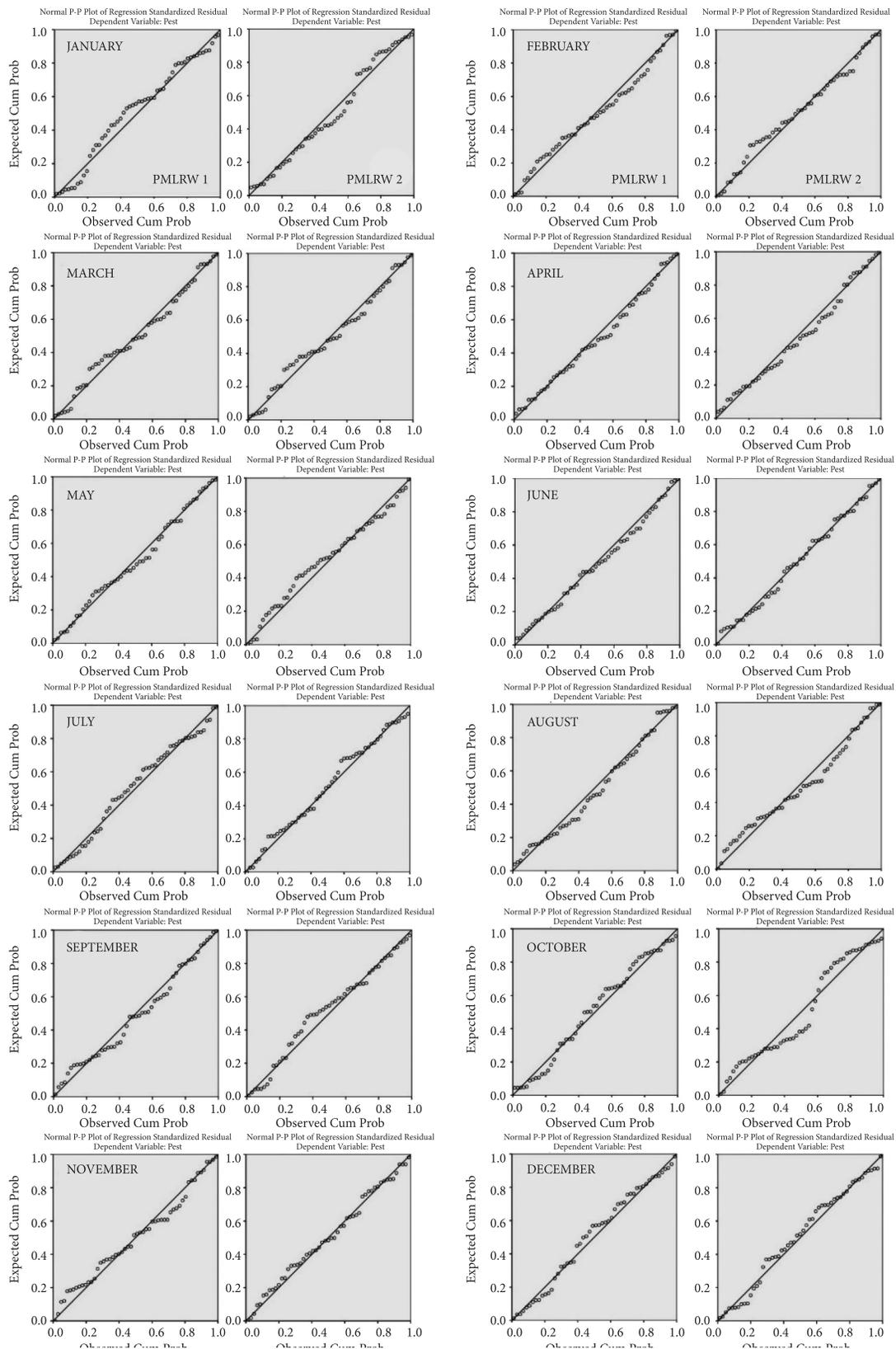
Figure S2. Normal P-P plots for January to December of PMLRW 1 (Model 1) and PMLRW 2 Model 2).
Left plot for Model 1 and right for Model 2 in each column
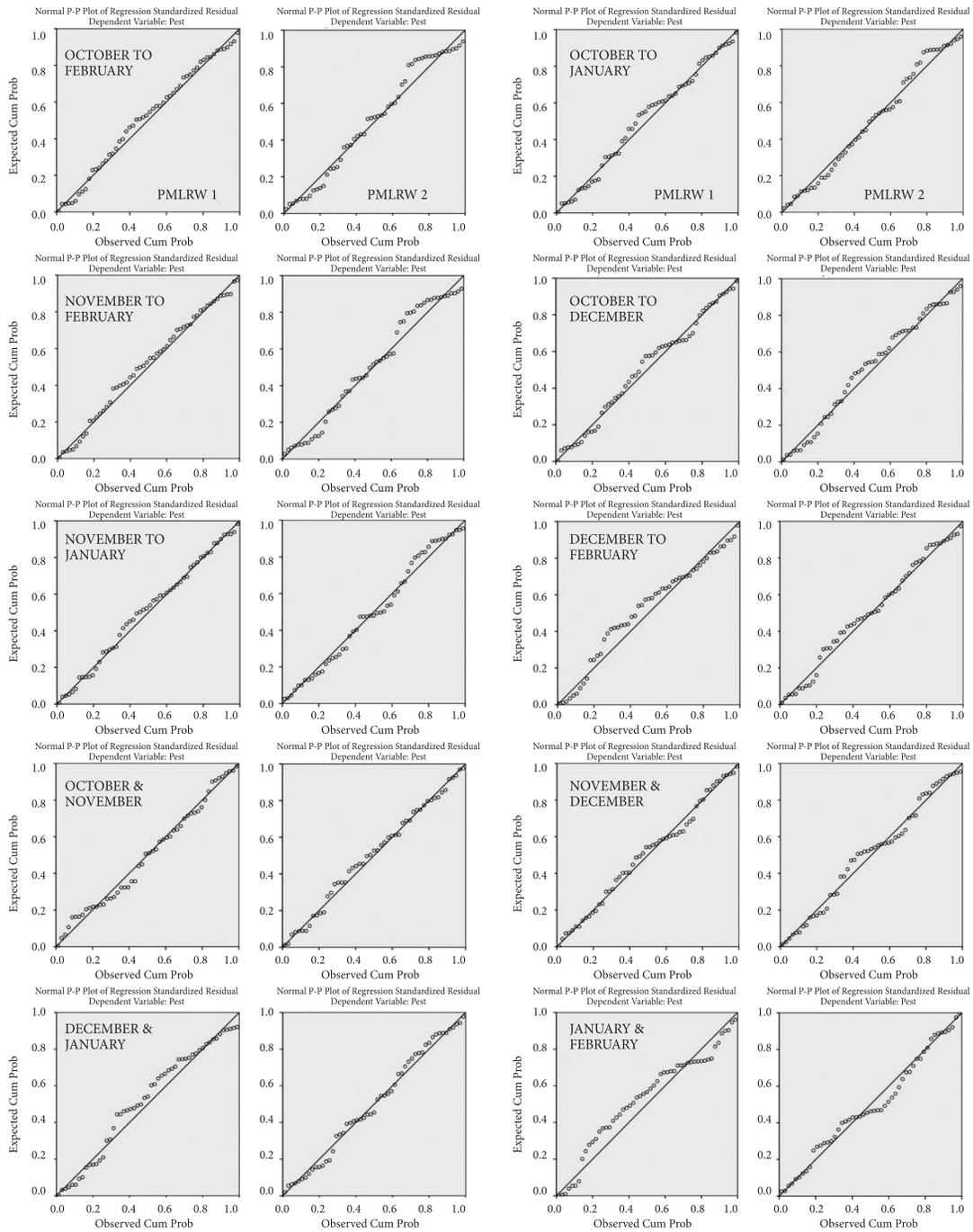
Figure S3. Normal P-P plots for all combinations of months for PMLRW 1 (Model 1) and PMLRW 2 (Model 2).
Left plots are for Model 1 while right for Model 2 in each column
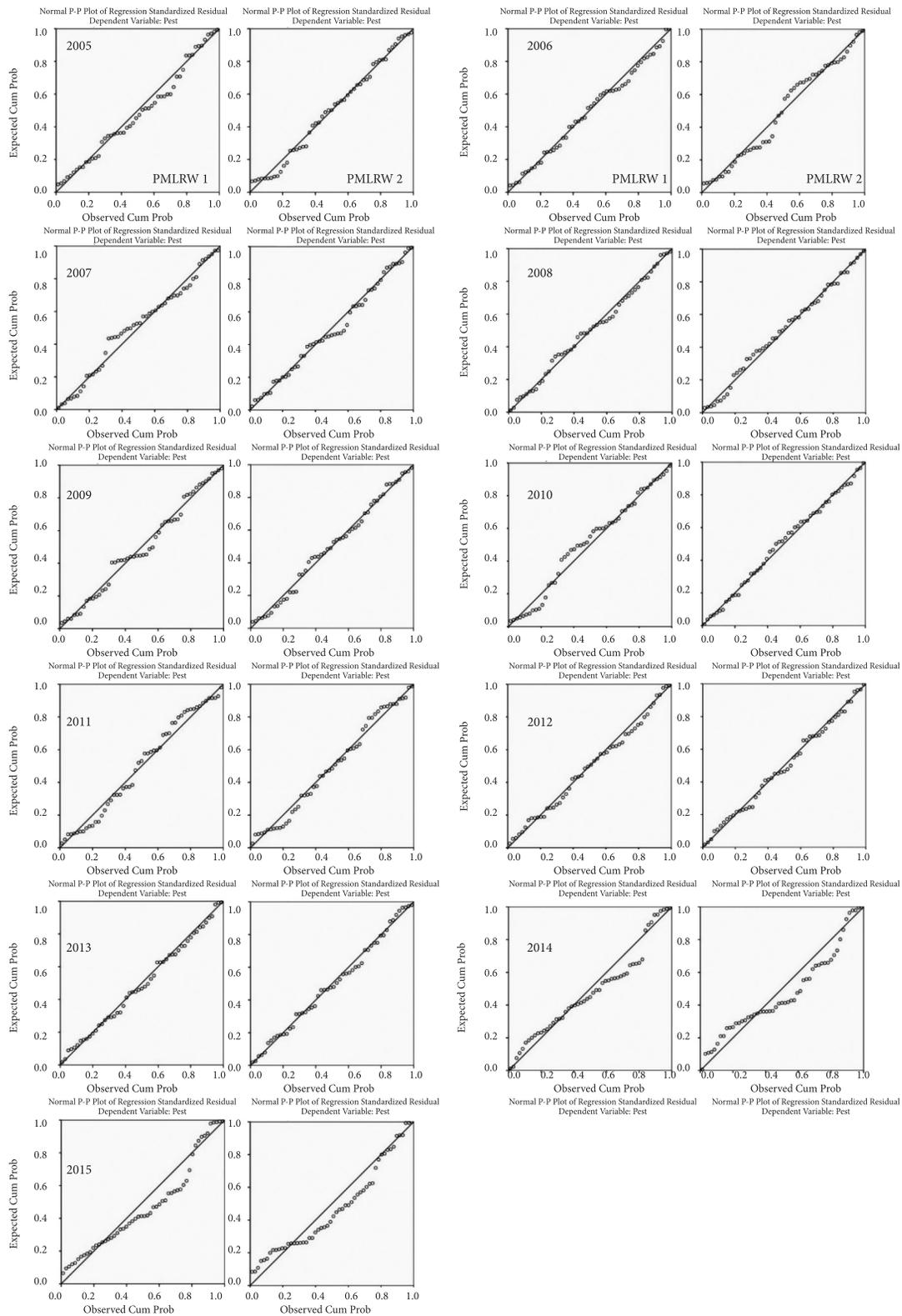
Figure S4. Normal P-P plots for all years for PMLRW 1 (Model 1) and PMLRW 2 (Model 2).
Left plots are for Model 1 while right for Model 2 in each column B