Remote Sensing of Environment 239 (2020) 111606

Contents lists available at ScienceDirect

Remote Sensing of Environment

journal homepage: www.elsevier.com/locate/rse

RF-MEP: A novel Random Forest method for merging gridded precipitation products and ground-based measurements

Oscar M. Baez-Villanueva[a,b,*], Mauricio Zambrano-Bigiarini[c,d], Hylke E. Beck[e], Ian McNamara[a], Lars Ribbe[a], Alexandra Nauditt[a], Christian Birkel[f,g], Koen Verbist[h], Juan Diego Giraldo-Osorio[i], Nguyen Xuan Thinh[b]

[a] Institute for Technology and Resources Management in the Tropics and Subtropics (ITT), TH Köln, Cologne, Germany
[b] Faculty of Spatial Planning, TU Dortmund University, Dortmund, Germany
[c] Department of Civil Engineering, Universidad de la Frontera, Temuco, Chile
[d] Center for Climate and Resilience Research, Universidad de Chile, Santiago, Chile
[e] Department of Civil and Environmental Engineering, Princeton University, Princeton, USA
[f] Geography Department, University of Costa Rica, San José, Costa Rica
[g] Northern Rivers Institute, University of Aberdeen, Aberdeen, UK
[h] UNESCO International Hydrological Programme, Paris, France
[i] Pontificia Universidad Javeriana, Bogotá, Colombia

ARTICLE INFO

Edited by Menghua Wang

Keywords:
Bias correction
Merging
Precipitation
Precipitation products
Random Forest
RF-MEP

ABSTRACT

The accurate representation of spatio-temporal patterns of precipitation is an essential input for numerous environmental applications. However, the estimation of precipitation patterns derived solely from rain gauges is subject to large uncertainties. We present the Random Forest based MErging Procedure (RF-MEP), which combines information from ground-based measurements, state-of-the-art precipitation products, and topography-related features to improve the representation of the spatio-temporal distribution of precipitation, especially in data-scarce regions. RF-MEP is applied over Chile for 2000—2016, using daily measurements from 258 rain gauges for model training and 111 stations for validation. Two merged datasets were computed: RF-MEP$_{3P}$ (based on PERSIANN-CDR, ERA-Interim, and CHIRPSv2) and RF-MEP$_{5P}$ (which additionally includes CMORPHv1 and TRMM 3B42v7). The performances of the two merged products and those used in their computation were compared against MSWEPv2.2, which is a state-of-the-art global merged product. A validation using ground-based measurements was applied at different temporal scales using both continuous and categorical indices of performance. RF-MEP$_{3P}$ and RF-MEP$_{5P}$ outperformed all the precipitation datasets used in their computation, the products derived using other merging techniques, and generally outperformed MSWEPv2.2. The merged *P* products showed improvements in the linear correlation, bias, and variability of precipitation at different temporal scales, as well as in the probability of detection, the false alarm ratio, the frequency bias, and the critical success index for different precipitation intensities. RF-MEP performed well even when the training dataset was reduced to 10% of the available rain gauges. Our results suggest that RF-MEP could be successfully applied to any other region and to correct other climatological variables, assuming that ground-based data are available. An R package to implement RF-MEP is freely available online at https://github.com/hzambran/RFmerge.

1. Introduction

Precipitation (*P*) is a key parameter in the hydrological cycle and an accurate estimation of its spatio-temporal variability is therefore crucial for numerous hydrological, agricultural, and ecological purposes. *P* is commonly measured with rain gauge stations, with a high accuracy at specific locations (Villarini et al., 2008). If only ground-based measurements are used, the accuracy of the representation of spatial *P*

* Corresponding author.
E-mail addresses: obaezvil@th-koeln.de (O.M. Baez-Villanueva), mauricio.zambrano@ufrontera.cl (M. Zambrano-Bigiarini), hylke.beck@gmail.com (H.E. Beck), ian.mcnamara@th-koeln.de (I. McNamara), lars.ribbe@th-koeln.de (L. Ribbe), alexandra.nauditt@th-koeln.de (A. Nauditt), christian.birkel@ucr.ac.cr (C. Birkel), k.verbist@unesco.org (K. Verbist), j.giraldoo@javeriana.edu.co (J.D. Giraldo-Osorio), nguyen.thinh@tu-dortmund.de (N. Xuan Thinh).

https://doi.org/10.1016/j.rse.2019.111606
Received 30 March 2019; Received in revised form 10 October 2019; Accepted 11 December 2019
Available online 02 January 2020

Table 1
Studies that have applied merging procedures to improve the spatio-temporal characterisation of *P* at different temporal scales. Those marked with a star have derived merged *P* estimates at daily or sub-daily temporal resolution

Study	Merging method(s)	Region	Product(s)	Number of stations(training / validation)	Main results
Li and Shao (2010)★	Nonparametric kernel smoothing	Australia	TRMM 3B42v6	5007 (90%) / 556 (10%)	TRMM 3B42v6 had an RMSE of 6.04 mm, mean error (ME) of −0.19 mm, and coefficient of efficiency of 0.29. The merged product had an RMSE of 3.43 mm, ME of 0.05 mm, and CE of 0.74
Rozante et al. (2010)★	Barnes objective analysis method	South America	TRMM 3B42RT	~1350 (90%) / ~150 (10%)	The 5-day RMSE for TRMM 3B42RT ranged from ~4 to ~22 mm, and the POD ranged from ~0.63 (no-rain) to ~0.2 (heavy rain). The RMSE of the merged product ranged from ~2 to ~18 mm, while the POD ranged from ~0.75 (no-rain) to ~0.38 (heavy rain)
Xie and Xiong (2011)★	Optimal interpolation	China (validation over South Korea)	CMORPH	2400 (China); 600 (96%) / 28 (4%) (over South Korea)	Bias corrected CMORPH correlations ranged from 0.65 to 0.75, while the derived product correlations ranged from 0.70 to 0.98
Gebregiorgis and Hossain (2011)★	Linear weights based on hydrologic model predictability	Mississippi River Basin	TRMM 3B42RT, CMORPH, and PERSIANN-CCS	Calibration and validation through hydrological modelling	Validation over 6 catchments within the Mississippi River Basin. The Nash-Sutcliffe efficiency (NSE) using the products ranged from −304 to 0.6, while the NSE using the merged product derived using runoff weights ranged from −0.3 to 0.6
Woldemeskel et al. (2013)	Linearised weighting procedure	Australia	TRMM 3B42v6	207 (90%); 184 (80%) / 23 (10%); 46 (20%, representing 1 of 5 regions)	The RMSE of TRMM 3B42v6 was 27.8 mm and the ME −3.42 mm. The RMSEs of the merged product were 23.73 mm and 28.24 mm, and the MEs were −2.42 mm and −2.65 mm for 10% and 20% of the stations, respectively
Shen et al. (2014)★	Arithmetic mean and inverse-error-square weighting methods	Tibetan Plateau	CMORPH, PERSIANN, NRL, TRMM 3B42v7 and 3B42RT	~330 stations / no information about stations used for validation	Summer period: merged products showed slight improvements on *P* products. Winter period: TRMM 3B42v7 outperformed all merged products
Nie et al. (2015)★	Optimal interpolation	China	CMORPH and NCEP	1920 (80%) / 480 (20%)	The RMSEs of the products ranged from 6.05 to 6.68 mm; *r* from 0.53 to 0.61, and bias from −0.86 to 0.48. The RMSE of the merged product was 4.38 mm, *r* of 0.78, and bias of −0.02. POD ranged from 0.9 (no-rain) to 0.5 (heavy-rain) and FAR from ~0.27 to ~0.35, respectively
Fu et al. (2016)	Bayesian model averaging	China	MERRA, Princeton, ERA-Interim, CMAP, NCEP, GPCPv2.2, and GPCCv6	~378 (50%) / ~377 (50%)	The RMSEs of the products ranged from ~105 to ~265 mm and the R^2 from ~0.38 to ~0.81, while the merged product showed a RMSE of ~85 mm and an R^2 of ~0.89
Manz et al. (2016)	Linear modelling, residual IDW, and Kriging-based methods	Tropical Andes	TRMM 2A25	722 / 1 (Leave one-out cross validation)	The RMSE of the product ranged from ~70 to ~95 mm and the relative bias (RB) from -37% to -15%. Ordinary Kriging (OK) performed best with a range of RMSE from ~45 to ~65 mm and relative bias from ~-3% to ~3%
Verdin et al. (2016)	OK and k-nearest neighbour local polynomials	Central America, Colombia and northwestern Venezuela	CHIRP	Not specified / Leave one-out cross validation	The RMSE from CHIRP ranged from 67.6 to 200.1 mm and the bias from -49.0% to 25.7%. The OK RMSE ranged from 60.7 to 137.0 mm and the bias from -0.1% to 0.8%
Shi et al. (2017)★	Merging weights based on the effective influence radius of rain gauges	Beijing, China	CMORPH	115 stations / validation through hydrological modelling	Streamflow simulation: NSE from 0.38 (CMORPH) to 0.66 (merged product)
Yang et al. (2017)★	Inverse-root-mean-square-error weighting	Chile	PERSIANN-CCS	414 (90%) / 42 (10%)	Performances reported according to 3 geographic regions. The RMSE from PERSIANN-CCS ranged from 3.08 to 7.35 mm; *r* from 0.66 to 0.70; and bias from -3.28 to 1.11. The RMSE of the merged product ranged from 0.27 to 2.69 mm; *r* from 0.95 to 0.97; and bias from −0.04 to 0.02
Ma et al. (2018)★	Bayesian model averaging (BMA)	Tibetan Plateau	TRMM 3B42RT, TRMM 3B42v7, CMORPH, and PERSIANN-CDR	200 (93%) / 15 (7%)	The RMSE of the products ranged from 7.0 to 14.1 mm; RB from 72.6 to 130.0%; and the POD ranged from 0.69 (no-rain) to 0.78 (heavy-rain). The merged product presented an RMSE of 6.77 mm; an RB of 70.83%; and an overall POD of 0.90
Beck et al. (2019)★	Weighted averaging with CDF matching	Global (evaluated over the CONUS)	CMORPH, ERA-Interim, GPCC FDR, GridSat, GSMaP, JRA-55, TRMM 3B42RT, and WorldClim	76,747 *P* stations and 13,762 streamflow stations / Stage IV dataset	Validated using the NCEP Stage IV dataset as the ground-truth and compared to MSWEPv1, CMORPH, ERA-Interim, and MERRA-2. The modified Kling-Gupta efficiency (KGE') of the products ranged from 0.35 to 0.53. The KGE' of MSWEPv2.2 was 0.70

patterns relies on the density and configuration of the gauge network (Adhikary et al., 2015; Borga and Vizzaccaro, 1997; Chen et al., 2008; Garcia et al., 2008; Goudenhoofdt and Delobbe, 2009; Villarini and Krajewski, 2008). In particular, a high network density is of most importance to capture the spatial distribution of convective events (Garcia et al., 2008).

In many developing countries the network of rain gauges is sparsely distributed; therefore, the use of only ground-based measurements to estimate the spatial distribution of *P* is subject to large uncertainties (Woldemeskel et al., 2013). Elevation must be considered because of the important role it plays in the *P* process. In general, higher elevation causes more *P* (Jaagus et al., 2010), an effect that can be extremely pronounced even over small elevation changes. For example, Bergeron (1960) reported that precipitation rates over small hills were twice the value of the lower areas in a flat region of 30 km^2, with approximately 50 m elevation difference. In regions with complex topography, *P* is typically under-represented at higher elevations because most rain gauges are located in lowlands due to accessibility and economical considerations (Derin and Yilmaz, 2014).

Satellite and reanalysis-based *P* estimates (hereafter *P* products) provide an unprecedented opportunity to estimate the spatio-temporal distribution of *P* in regions with a sparse network of rain gauge stations. However, the evaluation of these products has shown that multiple sources of errors are still present (e.g., false detection, systematic, and random errors) and that these products tend to perform worse at shorter time scales (e.g., daily and sub-daily) than at longer time scales (e.g., monthly, seasonal, and annual), making their application difficult for hydrological modelling (Maggioni and Massari, 2018). Therefore, a need remains to improve the spatio-temporal distribution of *P* by combining different data sources such as *P* products and ground-based information (Xie and Xiong, 2011).

Several approaches have been implemented to derive gridded *P* and other climatological variables using point-based information and gridded products. These include optimal interpolation (OI) (Xie and Xiong, 2011), the linearised weighting procedure (Woldemeskel et al., 2013), non-parametric kernel smoothing (Li and Shao, 2010), Kriging-based methods (Seo et al., 1990; Grimes et al., 1999; 1990; Verdin et al., 2016), conditional merging (Sinclair and Pegram, 2005), partial thin plate splines (Hutchinson, 1995; McKenney et al., 2006; McVicar et al., 2007), among others. Table 1 lists merging studies used to improve the characterisation of *P*, with a more detailed description of the steps employed in each method included in the Table A1 from Appendix A.

Despite the improvements in the spatio-temporal representation of *P* achieved by these methods, many studies only merge the ground observations with a single *P* product (e.g., Li and Shao, 2010, Rozante et al., 2010, Shi et al., 2017, Verdin et al., 2016, Xie et al., 2017, Yang et al., 2017). Therefore, valuable information that is better captured by other products is not considered. Averaging *P* products (e.g., Shen et al., 2014) has negative effects in the detection of *P* intensities at daily temporal scale. The assumption of a Gaussian distribution is invalid for daily scales; therefore, the daily *P* data must be first transformed when using Bayesian model averaging Ma et al. (2018) or Kriging-based approaches. Furthermore, these merging methods are generally complex and difficult to implement.

Random Forest (RF; Biau and Scornet, 2016, Breiman, 2001, Prasad et al., 2006) is an ensemble learning method that can be used for supervised classification and regression tasks by constructing numerous decision trees using the relationship between independent and dependent variables. This technique is recognised for being accurate and able to deal with small sample sizes and high-dimensional feature spaces (Biau and Scornet, 2016). RF also performs well even when some explanatory variables do not add information to the prediction and when several covariates are used, mainly because it does not produce biased estimates or lead to overfitting (Biau and Scornet, 2016; Díaz-Uriarte and Alvarez de Andrés, 2006; Hengl et al., 2018). Although RF is a non-spatial technique, it can indirectly consider geographical covariates (e.g., coordinates, Euclidean distances to sampling locations, or downslope distances) and process-based covariates (e.g., elevation, rate of elevation change, or aspect).

Recently, Hengl et al. (2018) compared RF and several Kriging-based methods to evaluate whether RF was suitable for deriving spatial predictions of daily *P*. Although the performances of both methods were similar, they described several advantages in applying RF: *i)* there is no need to define an initial variogram; *ii)* the trend model is built automatically; *iii)* there is no need to define a search radius; *iv)* there are built-in protections against overfitting; and *v)* the method shows which individual observations and parameters are most influential. Therefore, RF is identified as an appropriate technique for merging *P* products with ground-based information, especially because different *P* products exhibit distinct performances and errors (e.g., under/overestimation, correlation with ground-based measurements, or detection of *P* events) depending on the region (Baez-Villanueva et al., 2018; Maggioni and Massari, 2018; Zambrano-Bigiarini et al., 2017).

In this study, the RF-based MErging Procedure (RF-MEP) is presented with the aim of improving the characterisation of the spatio-temporal distribution of *P* in data-scarce regions at any temporal scale. RF-MEP takes advantage of combining information from different *P* products, topography-related datasets, and *P* time series from rain gauges.

2. RF-MEP

RF-MEP is based on three key assumptions: *i*) *P* measurements from rain gauge stations are accurate at the point scale; *ii*) *P* products are generally biased but contain useful information about the spatio-temporal patterns of *P*; and *iii*) the combination of different *P* products and rain gauge data can provide a better representation of the spatio-temporal variability of *P* than any single product.

RF-MEP uses RF to predict the spatial distribution of *P* by merging information from different gridded products (known as covariates) and quality-controlled ground-based information at a selected temporal scale (e.g., daily, monthly, or annual). Individual predictions are generated from a user-defined number of decision trees based on bootstrap samples using the covariates as predictors. The final prediction is calculated as the average of the individual predictions (Biau and Scornet, 2016; Breiman, 2001; Hengl et al., 2018; Prasad et al., 2006; Roy and Larocque, 2012). Fig. 1 summarises the four steps involved in this method.

2.1. Data acquisition

First, the selected covariates and ground-based measurements are acquired. The spatial covariates are: *i*) the selected *P* products, and *ii*) topography-related datasets such as digital elevation model (DEM), aspect, rate of elevation change, or slope, which are used to account for the *P* gradient related to elevation (not to be mistaken with altitude, see McVicar and Körner, 2013). The ground-based measurements are quality controlled and checked for homogeneity.

2.2. Data processing

The selected rain gauge stations are divided into two groups: a training set (to train the RF model) and a validation set (to assess the performance of the merged product). The selected *P* products and topography-related datasets are resampled to a selected spatial resolution to ensure identical raster geometry (spatial resolution, spatial extent, and origin).

The traditional RF algorithm ignores sampling locations which could lead to sub-optimal predictions (Hengl et al., 2018); therefore, covariates that account for geographical proximity are incorporated. The use of only geographical coordinates as spatial predictors can cause

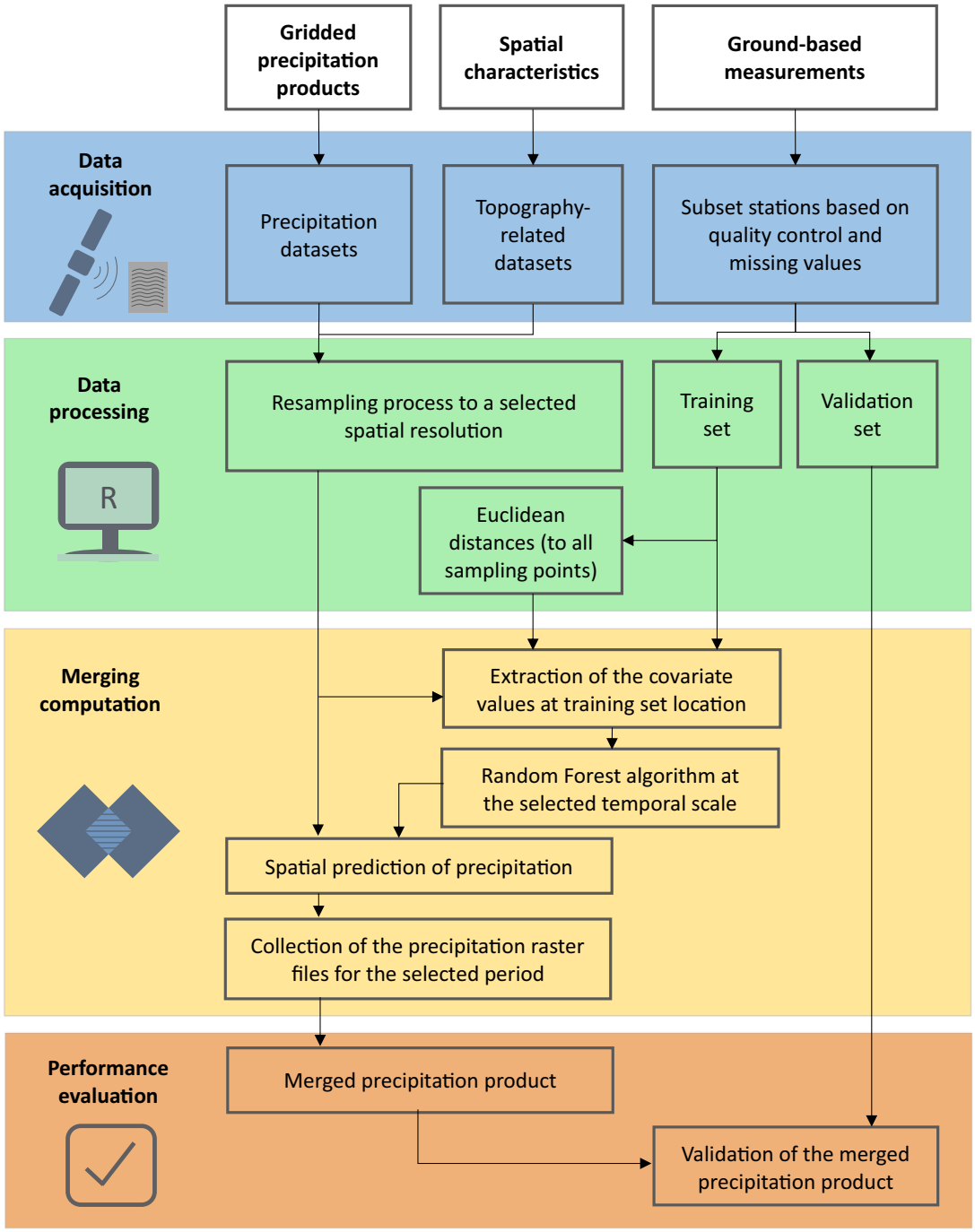

Fig. 1. Flow chart summarising RF-MEP, which is used to derive a better representation of the spatio-temporal distribution of *P* from the combination of *P* products, topography-related datasets, and ground-based data.

unnatural surfaces in the merged product (Behrens et al., 2018; Hengl et al., 2018). Instead, RF-MEP uses gridded layers of Euclidean distances from each rain gauge in the training set to the centroid of all the grid-cells in the selected study area.

2.3. Merging procedure

For each time step a single RF regression model is derived to compute a single *P* prediction at the desired temporal resolution. The RF model is trained using the ground-based observations in the training set as the dependent variable, while the grid-cell values of the selected covariates at the corresponding locations are used as predictors. To improve the accuracy and stability, and to reduce the variance and overfitting of the RF predictions, they are generated as an ensemble estimate from the numerous decision trees (Díaz-Uriarte and Alvarez de Andrés, 2006; Hengl et al., 2018) as observed in Eq. (1):

$$\hat{\theta}^{B}(x) = \frac{1}{B}\sum_{b=1}^{B} t_b^*(x) \quad (1)$$

where $\hat{\theta}^{B}$ is the final prediction; b is the individual bootstrap sample; B is the total number of trees; and t_b^* is the individual decision tree. This process is repeated for each time step, implying that the RF model will vary temporally. Fig. 2 illustrates an example of the merging procedure process using two *P* products, a digital surface model (DSM), three rain gauge stations, and the three correspondent Euclidean distance layers

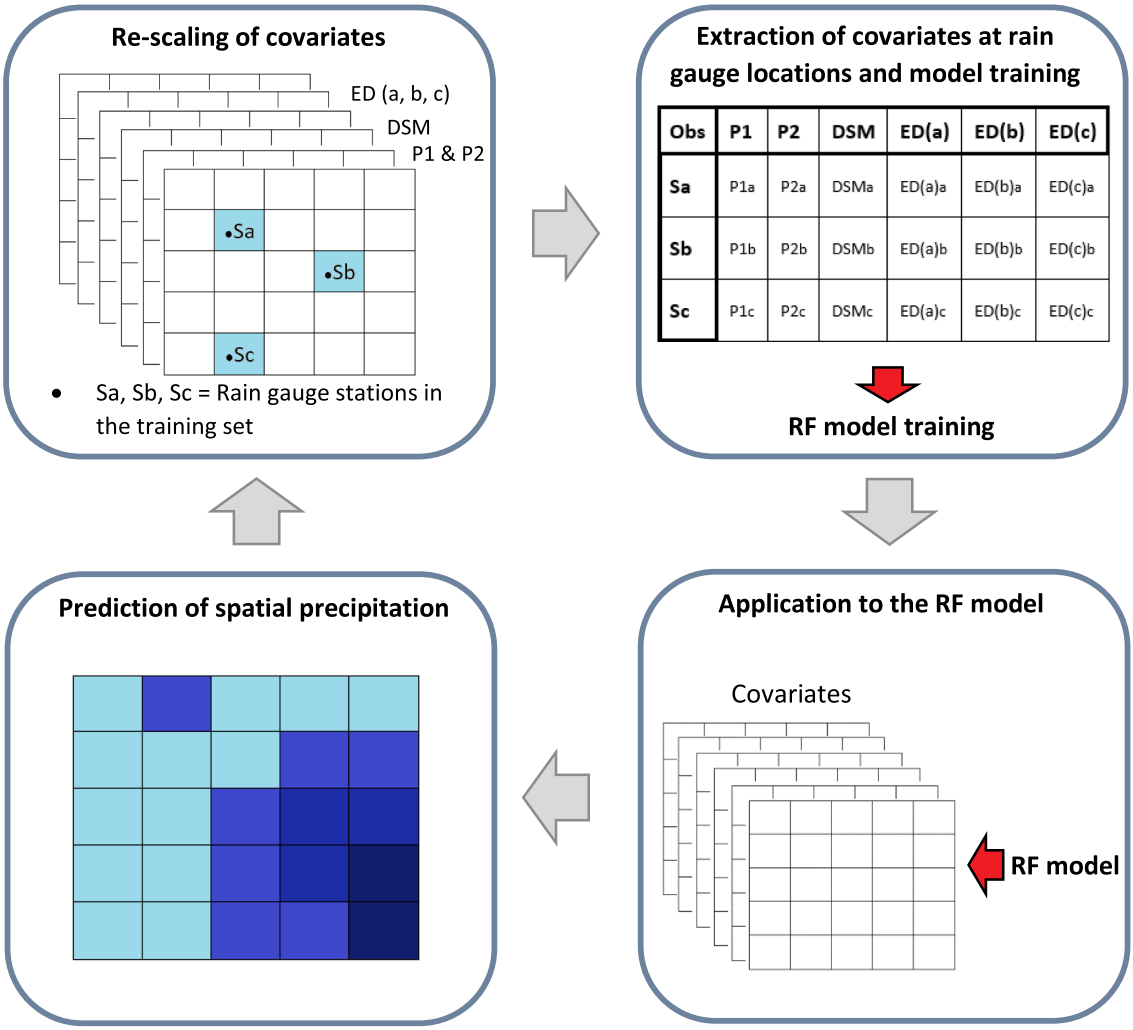

Fig. 2. Illustration of the merging procedure using two *P* products (P1 and P2), a DSM (to account for the topography-related datasets), three rain gauge stations (S_a, S_b, and S_c), and the three Euclidean distance layers ($ED_{(a)}$, $ED_{(b)}$, and $ED_{(c)}$).

($ED_{(a)}$, $ED_{(b)}$, and $ED_{(c)}$). An R package to implement RF-MEP is freely available online at https://github.com/hzambran/RFmerge.

2.4. Validation process

The validation set of rain gauge stations is used to assess the performance of the merged product using a point-to-pixel analysis, where the rain gauge station measurements are compared against the corresponding grid-cell values of the *P* products under the assumption that the rain gauge measurements are representative values at their respective grid-cells. However, this assumption may introduce bias in the comparison because: *i*) during winter, some rain gauges located at high elevation are not able to incorporate snow into the *P* measurement; and *ii*) during summer, a more dense network of rain gauges is required to capture the spatial patterns of small-scale convective events. Despite this, the point-to-pixel analysis is widely used to assess the performance of *P* products (e.g., Baez-Villanueva et al., 2018, Dinku et al., 2007, Gao and Liu, 2013, Hirpa et al., 2010, Li et al., 2013, Thiemig et al., 2012, Zambrano-Bigiarini et al., 2017). Among the plethora of indices available to assess the performance of *P* products, we selected the modified Kling-Gupta efficiency (KGE'; Gupta et al., 2009, Kling et al., 2012) over the traditional root mean squared error (RMSE) because the latter assigns disproportional weights to different *P* intensities at the daily scale (Baez-Villanueva et al., 2018). This is due to the high skewness of the precipitation distribution at the daily scale and the prevalence of temporal mismatches between estimated and observed precipitation peaks. The KGE' (Eq. 2) compares observed data with estimations, decomposing the total performance into three components: the linear correlation (r), the bias ratio (β), and the variability ratio (γ), presented in Eqs. (3), (4), and (5), respectively:

$$KGE' = 1 - \sqrt{(r-1)^2 + (\beta-1)^2 + (\gamma-1)^2} \quad (2)$$

$$r = \frac{\sum_{i=1}^{n}(O_i - \bar{O})(S_i - \bar{S})}{\sqrt{\sum_{i=1}^{n}(O_i - \bar{O})^2}\sqrt{\sum_{i=1}^{n}(S_i - \bar{S})^2}} \quad (3)$$

$$\beta = \frac{\mu_s}{\mu_o} \quad (4)$$

$$\gamma = \frac{CV_s}{CV_o} = \frac{\sigma_s/\mu_s}{\sigma_o/\mu_o} \quad (5)$$

where n is the number of observations; O_i and S_i are the observed and simulated values of the corresponding *P* product at day i; and $\bar{O}$ and $\bar{S}$ are the arithmetic means of the observations and the *P* product, respectively. r measures the temporal *P* dynamics; β measures the total *P* volume compared to ground-based observations indicating the average tendency of the *P* products to underestimate ($\beta < 1$) or overestimate ($\beta > 1$); and γ measures the relative dispersion between the gridded product and the ground-based measurements (Gupta et al., 2009; Kling et al., 2012). The optimal value for the KGE' and all its components is

one. The KGE' is a useful evaluation index because: *i*) it does not assign disproportional weights to mismatches in high precipitation values (contrary to squared-difference indices; e.g., the RMSE); *ii*) it decomposes the total performance into three components, thus allowing a better understanding of the origin of mismatches (Baez-Villanueva et al., 2018; Zambrano-Bigiarini et al., 2017); and *iii*) it allows a fair comparison of regions with different mean annual *P*. The KGE' has been widely used in hydrological applications and to evaluate the performance of *P* products (e.g., Baez-Villanueva et al., 2018, Beck et al., 2016, 2017b, Chen et al., 2014, Lievens et al., 2015, Thiemig et al., 2013, Wang et al., 2018, Zambrano-Bigiarini et al., 2017).

To evaluate the performance of *P* products in capturing different *P* intensities we used several categorical indices of performance: the probability of detection (POD; Eq. (6)), frequency bias (*f*bias; Eq. (7)), false alarm ratio (FAR; Eq. (8)), and critical success index (CSI; Eq. (9)).

$$\mathrm{POD} = \frac{H}{H + M} \tag{6}$$

$$f\mathrm{bias} = \frac{H + F}{H + M} \tag{7}$$

$$\mathrm{FAR} = \frac{F}{H + F} \tag{8}$$

$$\mathrm{CSI} = [(POD)^{-1} + (1 - FAR)^{-1} - 1]^{-1} \tag{9}$$

where *H* indicates a hit (an event recorded by both the rain gauge and the *P* product); *M* indicates a miss (an event only identified by the rain gauge); and *F* indicates a false alarm (an event recorded only by the *P* product). The POD calculates how often the product correctly estimates the precipitation intensity observed at the rain gauge. The *f*bias compares the number of events identified by the *P* product to the number of events registered by the gauge station. If *f*bias > 1, the number of occurrences of the respective *P* intensity is overestimated by the product, while *f*bias < 1 indicates underestimation. The FAR measures the fraction of events that were not correctly identified by the *P* product. Finally, the CSI combines the POD and FAR to describe the overall ability of the products to correctly detect different *P* intensities. The POD, *f*bias, and CSI present their optimal value at one, while FAR presents it at zero.

3. Case study

The Chilean territory was selected as the case study to test the performance of the proposed RF-MEP due to the notable heterogeneity in topography, climate and land cover.

3.1. Study

Chile is a South American country with nearly 4300 km of latitudinal extension (from 17.5°S to 56.0°S) and an average longitudinal extension of around 180 km (from 76.0°W to 66.0°W). Chile is bounded to the north by Peru, to the east by Bolivia and Argentina, and to the west by the Pacific Ocean. The geography of the country is dominated by mountainous terrains, with an elevation profile ranging from 0 to 6891 m a.s.l. Morphologically, Chile exhibits four major geographical units distributed from east to west: the Andes Mountains, the Intermediate Depression, the Coastal Mountains, and the Coastal Plains (Valdés-Pineda et al., 2014). The four seasons of the southern hemisphere are present: autumn (MAM), winter (JJA), spring (SON), and summer (DJF). *P* tends to increase with latitude (in the southern direction) and elevation (Montecinos and Aceituno, 2003). The interannual variability of *P* is mostly related to the El Niño-Southern Oscillation (ENSO), which strongly impacts winter *P* patterns, generating positive anomalies during El Niño events and negative anomalies during La Niña events (Robertson et al., 2014; Verbist et al., 2010).

Fig. 3 shows the elevation (Jarvis et al., 2008), the Köppen-Geiger climate zones (Beck et al., 2018), and the most updated Chilean land cover classification (Zhao et al., 2016), dividing the country according to the five major macroclimatic zones defined in Zambrano-Bigiarini et al. (2017). A variety of climates are observed throughout Chile: arid and semi-arid climates in the north with extremely low *P* (≤ 50 mm yr^{-1}) and high temperatures; temperate climates in Central Chile; and humid climates in the southern regions, with *P* values reaching up to 5000 mm yr^{-1}. Furthermore, polar and tundra climates are observed in the highest elevations of the Andes Mountains. Land cover is characterised by barren land in the Far North, which transitions to forest in the Near North. Forest, grasslands, and croplands are present in Central Chile and the two southern regions, while grassland, forest, and snow/ice areas are predominantly observed in the Far South.

3.2. Datasets

3.2.1. Ground-based precipitation

Time series of ground-based daily *P* for 1900–2018 were downloaded from a database of 816 rain gauges from the Center of Climate and Resilience Research (CR2; http://www.cr2.cl/recursos_y_publicaciones/bases-de-datos/). These data are provided by Dirección General de Aguas (DGA) and Dirección Meteorológica de Chile (DMC), the Chilean water and meteorological agencies, respectively. In Chile, daily *P* is recorded at 08:00 local time (11:00–10:59 UTC).

3.2.2. SRTM-v4

We used the Shuttle Radar Topography Mission version 4 (SRTM-v4) DSM, which incorporates offsets due to vegetation height (Gallant et al., 2012), and has a reported vertical error of less than 16 m (Jarvis et al., 2008). We used the gap-filled SRTM-v4 product at a spatial resolution of 250 m.

3.2.3. Precipitation products

We selected six global or quasi-global state-of-the-art *P* products with at least 15 years of daily estimates (Table 2). These products were selected because: *i*) RF-MEP can be transferred to any selected study area using the same *P* products (or others) if ground-based data are available; and *ii*) the selected *P* products perform well in the study area (Baez-Villanueva et al., 2018; Zambrano-Bigiarini, 2018; Zambrano-Bigiarini et al., 2017).

The selected *P* products used in RF-MEP were: ERA-Interim (Dee et al., 2011); the Climate Hazards InfraRed Precipitation with Stations data version 2.0 (CHIRPSv2; Funk et al., 2015); the TRMM Multi-satellite Precipitation Analysis (TRMM 3B42v7; Huffman et al., 2010, 2007); the Precipitation Estimation from Remotely Sensed Information Using Artificial Neural Networks - Climate Data Record (PERSIANN-CDR; Ashouri et al., 2015, Sorooshian et al., 2000); and the Climate Prediction Center (CPC) Morphing technique version 1.0-BLD, gauge-satellite blended precipitation product (CMORPHv1; Joyce et al., 2004, Xie et al., 2017). The Multi-Source Weighted-Ensemble Precipitation (MSWEPv2.2; Beck et al., 2017a, 2019) was only used in the validation step as a benchmark product because: *i*) it is the first fully global *P* dataset derived by optimally merging a range of gauge, satellite, and reanalysis estimates (Beck et al., 2019); *ii*) it has shown more realistic spatial *P* patterns in mean, magnitude, and frequency than other state-of-the-art global precipitation products at the global scale (Beck et al., 2017b, 2019); *iii*) it uses the same rain gauge dataset within Chile; and *iv*) it recently outperformed other state-of-the art *P* products over Chile (Zambrano-Bigiarini, 2018). Detailed descriptions of the algorithms used by each *P* product can be found in their corresponding literature (see Table 2).

It is important to note that several *P* products use ground-based *P* data from the Global Precipitation Climatology Centre (GPCC; Schneider et al., 2008) to reduce bias (see Table 2). The number of operational GPCC rain gauge stations in Chile has fluctuated between seven and twenty over 1986–2018. This low density of GPCC stations

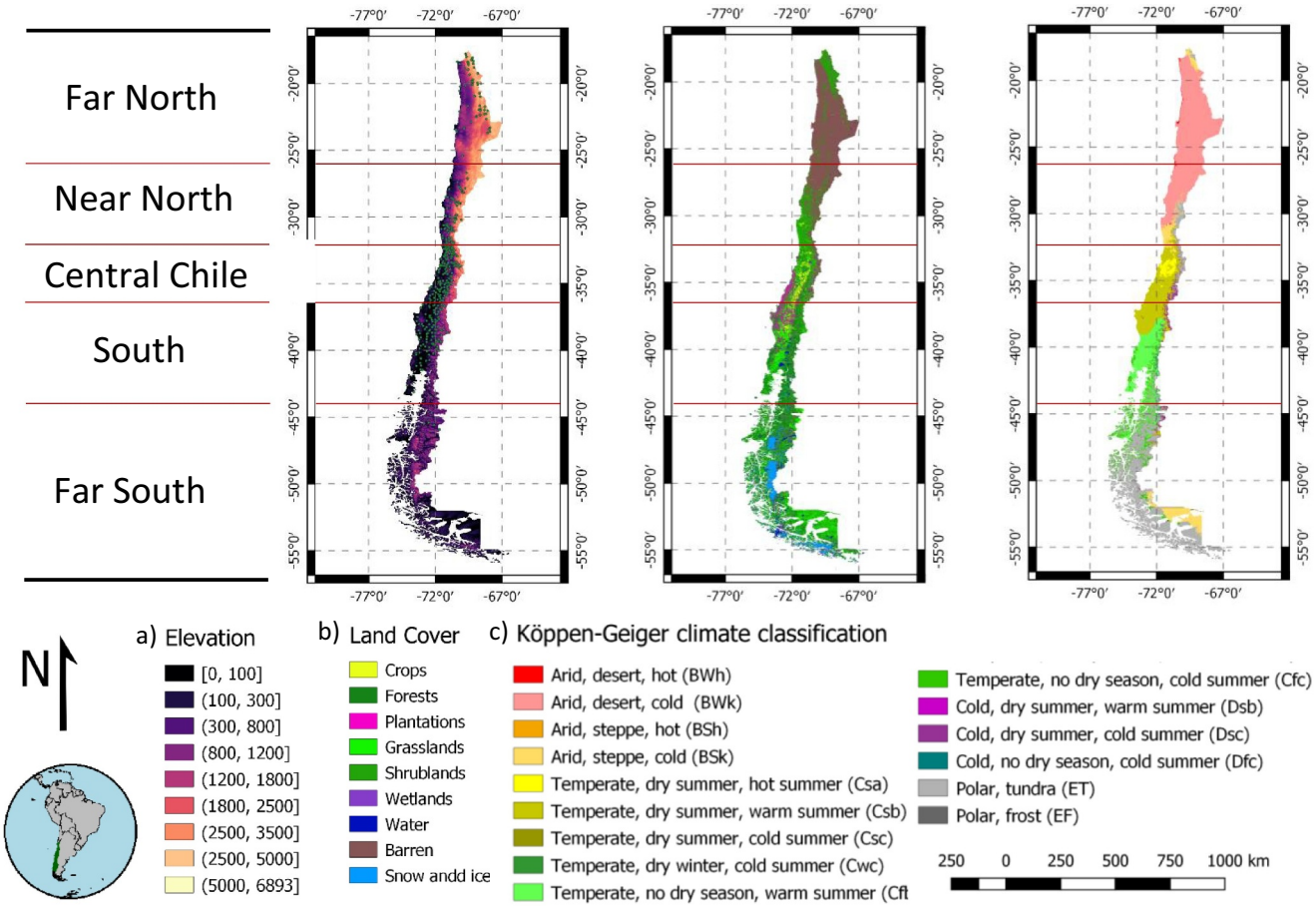

Fig. 3. Study area. (a) Elevation (Jarvis et al., 2008), including the rain gauge stations used in this case study; (b) land cover classification (Zhao et al., 2016); and (c) climate zones based on the Köppen-Geiger classification (Beck et al., 2018).

within Chile is clearly insufficient to adequately represent the spatio-temporal variability of *P* over the country.

3.2.4. Covariates

In addition to the selected *P* products and the DSM, other spatial covariates (slope, aspect, Köppen-Geiger climate classification, land cover type) were exhaustively evaluated using the KGE' and its components to ascertain whether an improvement could be obtained. Only the DSM was selected because the inclusion of the other covariates did not improve the performance of the final product.

3.3. Application of RF-MEP to the study area

RF-MEP was applied to the Chilean territory from 17.5° to 46.0°S for 2000–2016. The southern boundary was set due to the sparse network of gauge stations in the Far South. We used the R environment 3.5.0 (R Core Team, 2018) and the raster (Hijmans, 2018), hydroGOF (Zambrano-Bigiarini, 2017a), hydroTSM (Zambrano-Bigiarini, 2017b), GSIF (Hengl, 2019), and randomForest (Liaw and Wiener, 2002) R packages.

3.3.1. Data processing

All selected *P* products that are sub-daily (Table 2) were aggregated to the daily scale. MSWEPv2.2 was obtained at daily temporal scale because the 3-hourly version is not freely available. We downscaled PERSIANN-CDR, ERA-Interim, CMORPHv1 and TRMM 3B42v7 to the same spatial resolution as CHIRPSv2 (0.05°) using the nearest neighbour method (to avoid any improvements in the products performance prior to the merging procedure), while the DSM was upscaled from its original spatial resolution (250 m) to 0.05° using bilinear interpolation. The reason for resampling all the covariates to 0.05° (the highest spatial resolution of the selected *P* products) was to obtain a merged product that can be fairly compared to all selected *P* products.

We selected the 369 rain gauge stations that had < 5% of missing values and showed consistency when evaluated using the double-mass curve method to identify abnormalities comparing each station with the

Table 2
P products used in the case study.

Product	Spatial res.	Temporal res.	Period	Spatial coverage	Source(s)	Reference(s)
ERA-Interim	0.75°	3 hourly	1979–present	Global	Reanalysis	Dee et al. (2011)
CHIRPSv2★	0.05°	Daily	1981–present	50°N – 50°S	Satellite, gauge, and reanalysis	Funk et al. (2015)
TRMM 3B42v7★	0.25°	3 hourly	1998–present	50°N – 50°S	Satellite and gauge	Huffman et al. (2010, 2007)
PERSIANN-CDR★	0.25°	6 hourly	1983–2017 (April)	60°N – 60°S	Satellite and gauge	Ashouri et al. (2015), Sorooshian et al. (2000)
CMORPHv1★	0.25°	30 min	1998–present	60°N – 60°S	Satellite and gauge	Joyce et al. (2004), Xie et al. (2017)
MSWEPv2.2★	0.10°	3 hourly	1979–present	Global	Satellite, gauge, and reanalysis	Beck et al. (2017a, 2019)

★ Products that use GPCC data.

Table 3
Classification of *P* events in Chile based on daily intensity (*i*) according to Zambrano-Bigiarini et al. (2017).

Precipitation event	Intensity (*i*) in mm d^{-1}
No rain	[0, 1)
Light rain	[1, 5)
Moderate rain	[5, 20)
Heavy rain	[20, 40)
Violent rain	≥ 40

neighbouring stations, assuming homogeneity (Weiss and Wilson, 1953). The period 2000–2016 was chosen because of ground-based data availability over the period of record of the selected *P* products. A random sample of 70% of the selected rain gauge stations (258) were used as ground truth data to train the RF model (training set), while the remaining 30% of the stations (111) were used to assess the performance of the merged products (validation set). Past studies have typically selected 80% or more stations for training purposes (e.g., Li and Shao, 2010, Ma et al., 2018, Rozante et al., 2010, Woldemeskel et al., 2013, Yang et al., 2017); however, we selected 70% to be more thorough in the evaluation of the method. We computed the 258 layers of Euclidean distances using the GSIF R package (Hengl, 2019).

3.3.2. Merging procedure

Two merged *P* products were computed at the daily scale for 2000–2016. The first product (hereafter, RF-MEP_{3P}) used CHIRPSv2, PERSIANN-CDR, ERA-Interim, the DSM, and the 258 layers of Euclidean distances, while the second product (hereafter, RF-MEP_{5P}) added CMORPHv1 and TRMM 3B42v7 to the aforementioned covariates. The reason for computing two different merged products was to evaluate whether the addition of CMORPHv1 and TRMM 3B42v7, both of which have a shorter period of temporal coverage, would improve the final merged product. Although RF-MEP_{3P} and RF-MEP_{5P} were produced and compared over the same period (2000–2016), RF-MEP_{3P} can be generated over a longer period of record (1983–2016), while RF-MEP_{5P} can only be generated from 1998 onwards.

First, we obtained the values of the covariates at the grid-cell locations of the training set. Second, for each day, an RF model was trained using the ground-based *P* values as the dependent variable, and the respective values from the covariates as predictors. Third, the trained RF model was used with the gridded covariates to predict daily *P* values for each grid-cell of the study area. This process was repeated for each day for 2000–2016. RF regression models have three parameters to specify: *i*) the number of regression trees (set at 2000); *ii*) the number of variables randomly sampled at each decision split (set at one third of the number of covariates); and *iii*) the node size (i.e., the minimum number of observations per node; set at 5).

3.3.3. Performance evaluation

We evaluated the performance of both merged products, MSWEPv2.2, and the individual *P* products used as covariates, through a point-to-pixel analysis with the indices of performance described in Section 2.4, applied for the stations included in the validation set. The evaluation process was performed at multiple temporal scales: 3-day, monthly, annual, DJF, MAM, JJA, and SON.

Because no sub-daily measurements are available to transform the ground-based *P* dataset (see Section 3.2.1) to the 0:00–23:59 UTC daily period used by all the *P* products, we used 3-day accumulations as a proxy for evaluating daily performance. This approach reduces likely biases in the performance of the *P* products at this temporal scale by considering the influence of reporting times.

The categorical indices were evaluated using *P* intensities (Table 3; Zambrano-Bigiarini et al., 2017) recommended specifically for Chile.

Because the aim of RF-MEP is to improve the characterisation of *P* in data-scarce regions, we investigated the influence of the amount of rain gauge stations included in the training set. We computed the RF-MEP_{5P} product with varying percentages of rain gauge stations in the training set to evaluate the performance of RF-MEP under different data-scarcity scenarios. We computed the RF-MEP_{5P} product using 50%, 30%, and 10% of the stations, representing 184, 111, and 37 rain gauges, respectively.

To test the influence of the different spatial resolutions of the selected *P* products, we computed RF-MEP_{5P} at 0.05°, 0.10°, and 0.25°. For this purpose, all covariates were resampled to these spatial resolutions before the application of the merging procedure. Finally, we applied two additional merging methods to compare RF-MEP against established and proven precipitation merging procedures. We computed Kriging with external drift (KED) using ERA-Interim (the best-performing product used to derive RF-MEP_{5P}) and the one-outlier-removed (OOR) arithmetic mean described in Shen et al. (2014). For a detailed explanation of KED please refer to Ly et al. (2011), Oliver and Webster (2014), and Hengl et al. (2018). We also compared the RF-MEP_{5P} against MSWEPv2.2 because it is a state-of-the-art merged *P* product.

4. Results

4.1. Temporal assessment of the merged products

Fig. 4 plots the KGE' values at the seven assessed temporal scales for the existing and merged *P* products using the ground-based validation set. Both merged products (RF-MEP_{3P} and RF-MEP_{5P}) performed similarly well, with median KGE' values of 0.83, 084 and 0.78 at the 3-day, monthly, and annual scale, respectively. The *P* products used in the merging method presented median KGE' values between 0.20 and 0.60 at the 3-day scale, which increased to between 0.35 and 0.70 at the monthly and annual scales. Both merged products outperformed the *P* products used in the merging procedure at all temporal scales, demonstrating that the combination of *P* products and ground-based measurements generates a better representation of the spatio-temporal variability of *P*.

The merged products performed better than MSWEPv2.2 at all temporal scales except DJF (summer), where all *P* products showed a reduced performance and a greater dispersion in the KGE' values. This low performance in summer is the reason why the *P* products exhibit lower KGE' values at the annual scale compared to the monthly scale.

Fig. 5 shows boxplots with the individual KGE' components (r, β, and γ) at all temporal scales. Both merged products present a median r value of 0.94 at the 3-day temporal scale, which is consistent with the improvements in r obtained by Xie and Xiong (2011) and Yang et al. (2017). Of the existing *P* products, MSWEPv2.2 performed best with a median value of 0.89, highlighting the advantage of merging gauge, satellite, and reanalysis products. At all time scales, RF-MEP_{3P} and RF-MEP_{5P} performed considerably better than the products used in their computation. This demonstrates that the method is able to substantially improve the correlation of the *P* products for the Chilean case study.

Fig. 5b plots the performance of the β component of the KGE', showing that RF-MEP_{5P}, RF-MEP_{3P}, MSWEPv2.2, and CHIRPSv2 were close to exhibiting no bias. Both merged datasets present lower dispersion than MSWEPv2.2 and CHIRPSv2 for all temporal scales except DJF. This result shows that the evaluated products are generally biased but contain useful information that can be combined with ground-based measurements to derive improved *P* estimates. In DJF, both merged products presented a $\beta > 1$ and were outperformed by MSWEPv2.2.

Fig. 5c shows the γ component of the KGE', highlighting that all datasets underestimated the variability of *P* at all temporal scales. MSWEPv2.2 best represented the variability of the ground-based measurements, followed closely by both merged datasets. The high values of γ obtained for MSWEPv2.2 were expected because this product uses the same daily ground-based Chilean dataset in its computation and accounts for the difference in reporting times. Both merged products

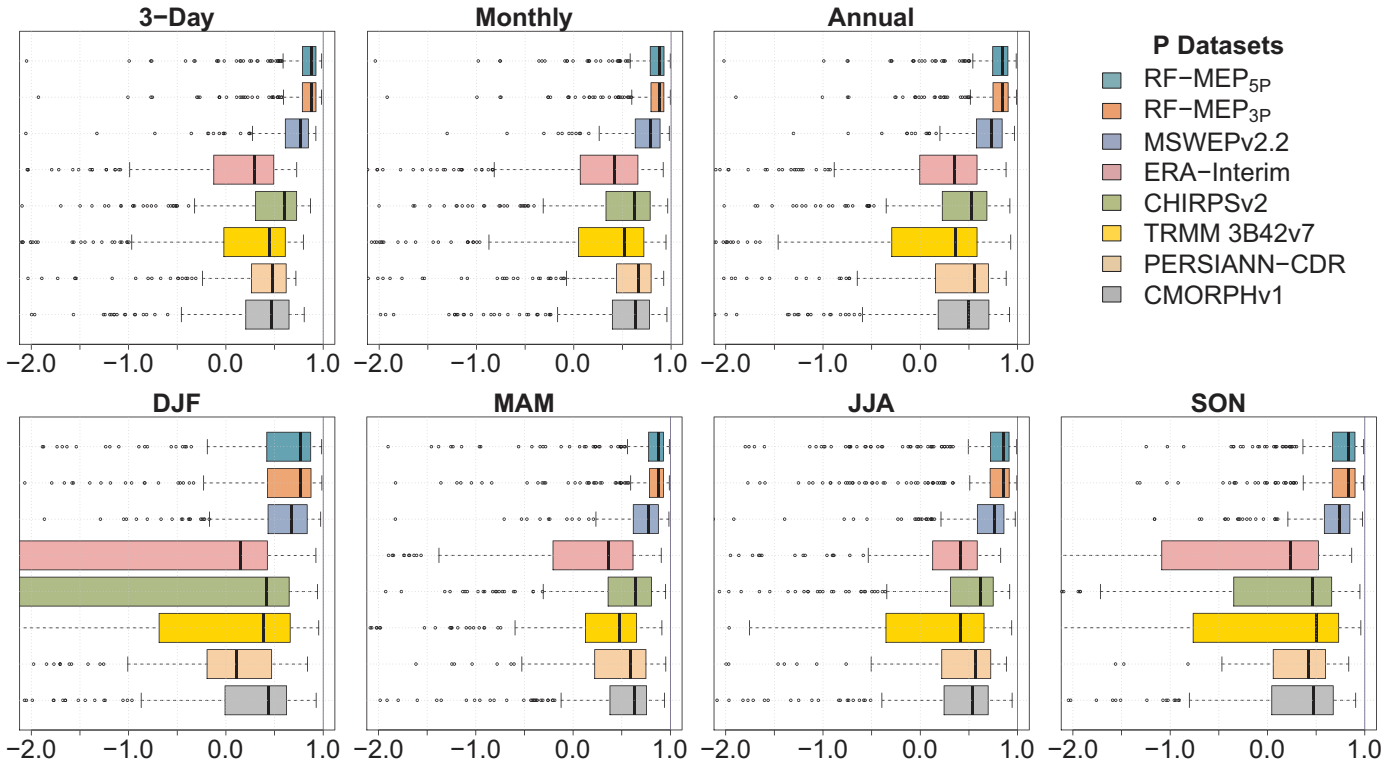

Fig. 4. KGE' values for all *P* products using the ground-based validation set. From left to right and top to bottom: 3-day, monthly, annual, DJF, MAM, JJA, and SON. The solid line represents the median value, the edges of the boxes represent the first and third quartiles, and the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. The blue line indicates the optimal value for the KGE'.

showed a reduced dispersion of the KGE' components at the 3-day, monthly, MAM, JJA, and SON scales; however, the dispersion at the annual scale increases due to the reduced performance in DJF.

4.2. Spatial assessment of the merged products

Fig. 6 summarises the KGE' of the 3-day *P* products over the four analysed macroclimatic zones, while Fig. 7 presents its spatial distribution. All products show median KGE' values lower than 0.5 and high dispersion in the Far North. These regions are classified as arid according to the Köppen-Geiger classification (see Fig. 3), demonstrating that the performance of the evaluated products over the arid regions of Chile remains low. MSWEPv2.2 and both merged products perform considerably better than the products used as covariates, highlighting the benefit of combining data from *P* products and ground-based measurements. The performance of all the products increased over Central Chile and South, where annual *P* volumes are much higher than in the Far North and Near North.

Fig. 7 shows that for both merged products, more than 80% of the stations in the validation set yielded KGE' values higher than 0.60. Both merged products performed best in the Near North, Central, and Southern Chile, with median KGE' values of 0.84, 0.86, and 0.81, respectively. However, in the Far North, MSWEPv2.2 performed the best (0.61), followed by RF-MEP$_{3P}$ (0.35) and RF-MEP$_{5P}$ (0.28). These results in the Far North show that the inclusion of more *P* products does not necessarily improve the median performance of the merged product; however, the inclusion of the additional two products reduced the dispersion in the KGE' values of RF-MEP$_{5P}$. Despite the poor performance of the *P* products used as covariates in the Far North, RF-MEP$_{3P}$ and RF-MEP$_{5P}$ were able to extract useful information from these products to obtain a better performance. RF-MEP$_{5P}$ and RF-MEP$_{3P}$ performed better in the high elevations of the Far North region compared to the low elevations (see Figs. 3 and 7). These high elevations correspond to the alpine tundra climate (ET), while the cold and arid desert climate (BWk) dominates the lower areas of the Far North, where the *P* datasets presented their worst performance. This suggests that arid climates present a great challenge for existing *P* products.

4.3. Assessment of precipitation intensities

Fig. 8 plots the median values of the four categorical indices for the five classes of daily *P* intensity described in Table 3. All datasets, with the exception of RF-MEP$_{3P}$ and RF-MEP$_{5P}$, obtained POD values lower than 0.45 for *P* events higher than 1 mm, while the no-rain events were well captured by all products. Similar results were observed for the FAR and CSI, where RF-MEP$_{3P}$ and RF-MEP$_{5P}$ presented the best performance of the evaluated products. FAR values were consistently the worst for the light rain intensities ([1, 5) mm d^{-1}), highlighting that the products remain unable to adequately capture low *P* values. The CSI presents the best performance for no-rain events followed by extreme events ($\geq$ 40 mm d^{-1}), as a result of the decreased FAR compared to the other *P* intensities.

Finally, the median values of the *f*bias showed that all *P* products overestimated the number of light rain ([1, 5) mm d^{-1}) and moderate rain events ([5, 20) mm d^{-1}). RF-MEP$_{3P}$ and RF-MEP$_{5P}$ performed the best in terms of *f*bias for the heavy rain events ([20, 40) mm d^{-1}), while MSWEPv2.2 performed the best for the other *P* intensities, followed by the merged products. All products underestimated the occurrence of violent rain events ($\geq$ 40 mm d^{-1}).

4.4. Impact of gauge density and spatial resolution of covariates

Fig. 9 shows the performance of RF-MEP$_{5P}$ with a varying number of stations used in the training set. The red line in the bottom left panel of Fig. 9 represents the median KGE' of the best-performing product used in the computation of RF-MEP$_{5P}$ (see Fig. 4), illustrating the

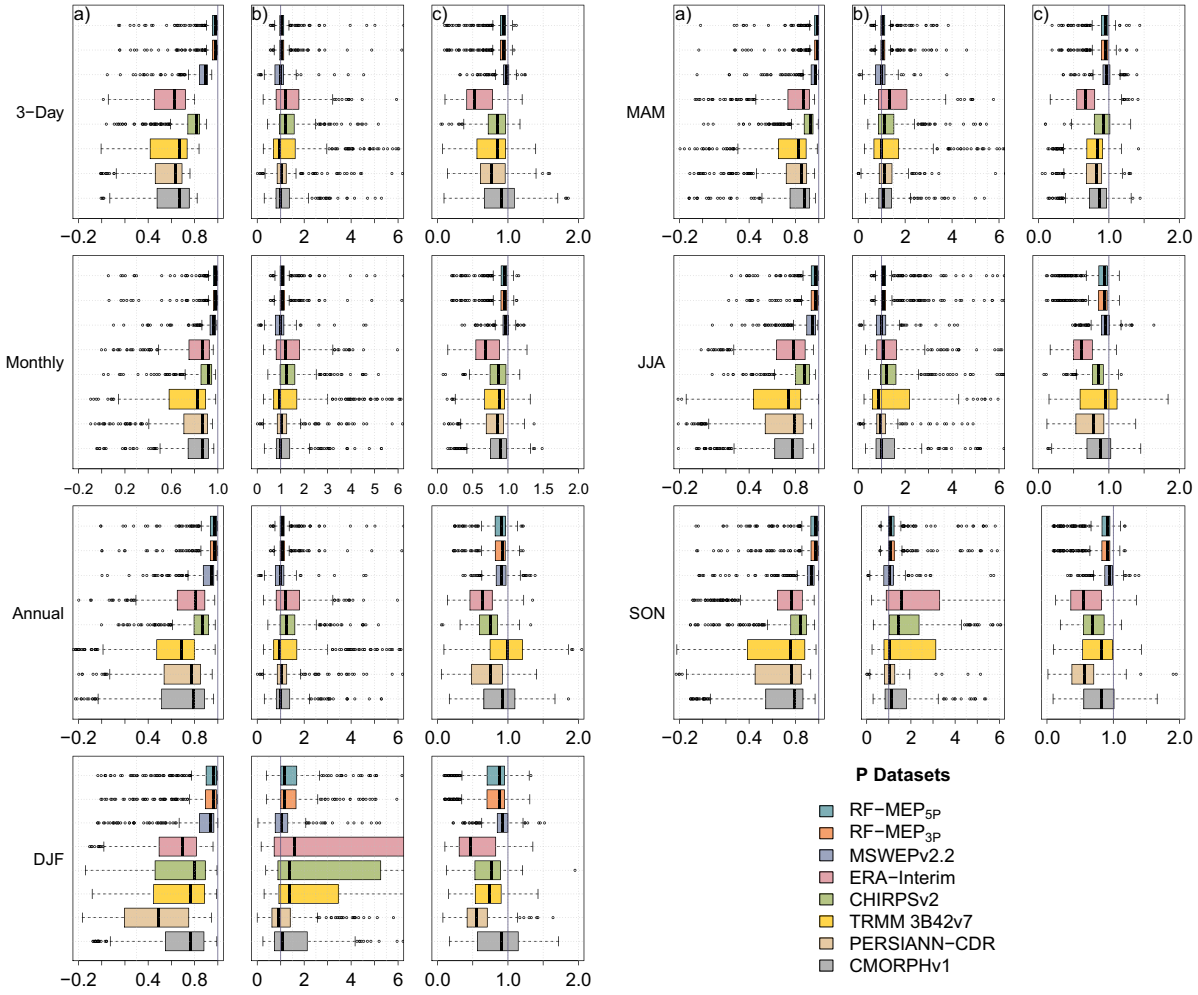

Fig. 5. The *r* (a), *β* (b), and γ (c) components of the KGE' for all *P* products using the ground-based validation dataset. The solid line represents the median value, the edges of the boxes represent the first and third quartiles, and the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. The blue line indicates the optimal value for each component.

improvement obtained even when only 10% (37) of stations are used in the training set. Also, Fig. 9 indicates that the inclusion of more stations improves the product performance in comparison to the best product available, which is consistent with other studies (Borga and Vizzaccaro, 1997; Chen et al., 2008; Goudenhoofdt and Delobbe, 2009). This suggests that the application of this method in other data-scarce regions is expected to improve the representation of *P*. The results of the CSI and *f*bias show that the RF-MEP$_{5P}$ increases the detection of different *P* intensities in comparison to the single *P* products (see Fig. 8). Similar to the KGE', there is a visible improvement in the detection of these events when more stations are used.

Fig. 10 plots the KGE' values of RF-MEP$_{5P}$ at all evaluated timescales for varying spatial resolutions of the covariates. It shows that resampling all the *P* products into a unified grid has a negligible impact on the performance of the final merged product.

The SRTM-v4 contains offsets in vegetated areas because the SRTM radar signal scatters from the woody structure within the canopy (Gallant et al., 2012). Although we did not remove the impacts of vegetation height to calculate a bare-earth DEM (~40 m over the South and Far South forests of Chile), we do not expect substantial changes because these elevation offsets become negligible at such a spatial resolution (0.05°).

4.5. Comparison between RF-MEP and different merging methods

Fig. 11 shows the performance of RF-MEP$_{5P}$ compared to KED, OOR arithmetic mean, and MSWEPv2.2. The performance of ERA-Interim is also plotted because it is the best-performing *P* product used in the merging procedure. RF-MEP$_{5P}$ showed the best performance at the 3-day temporal scale, followed by KED and MSWEPv2.2. The OOR arithmetic mean product shows the lowest KGE', γ, and *r*; however, it is able to accurately represent the total *P* volume at the 3-day scale. This product also shows the lowest performance when evaluated at different *P* intensities. Shen et al. (2014) concluded that the categorical performance of the OOR arithmetic mean product improved compared to the selected *P* products; however, they evaluated the categorical performance only for rain and no-rain events. The distribution of daily *P* is heavily skewed; and therefore, the performance of the product over different intensities can be masked by the no-rain events. As observed in the lower panel of Fig. 11, averaging different *P* products reduces the performance at all *P* intensities because all these products have errors in detection (i.e., the products may estimate different *P* intensities for a particular day). This analysis suggests that *P* products should not be averaged to attempt to improve daily *P* patterns.

KED performed similarly to RF-MEP$_{5P}$; however, RF-MEP$_{5P}$ showed

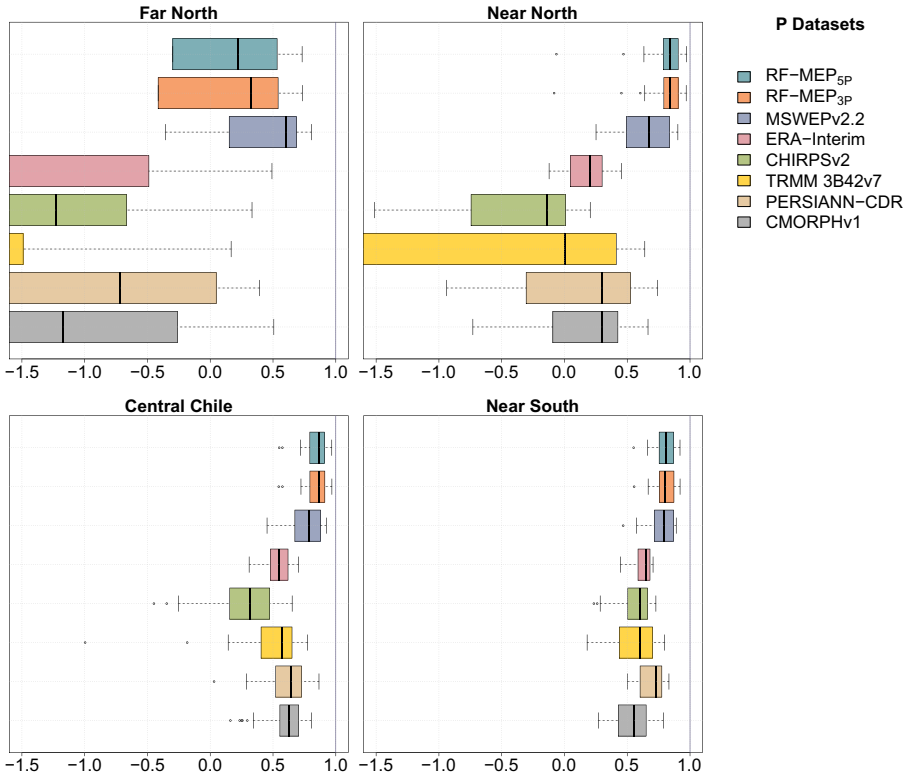

Fig. 6. 3-day KGE' values for the *P* products at the corresponding grid-cells of the validation set for the four analysed macroclimate zones: Far North, Near North, Central Chile, and South (see Fig. 3). The solid line represents the median value, the edges of the boxes represent the first and third quartiles, and the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. The vertical blue line indicates the optimal value for KGE'.

less dispersion in the KGE' and its components, suggesting that RF-MEP is a robust method to merge *P* products and ground-based data. Ly et al. (2011) obtained poor results when using KED with few sample points, which indicates that the performance of KED is highly influenced by the number of ground stations. Conversely, RF-MEP performed relatively well when the training set was dramatically reduced. The performance of RF-MEP$_{5P}$ is also the highest at monthly, annual and seasonal temporal scales, except in DJF where MSWEPv2.2 performs the best (see Fig. S1 in the supplementary material).

5. Discussion

5.1. Performance of the merged products

RF-MEP was applied at the daily temporal scale to derive two merged products (RF-MEP$_{5P}$ and RF-MEP$_{3P}$), which outperformed those

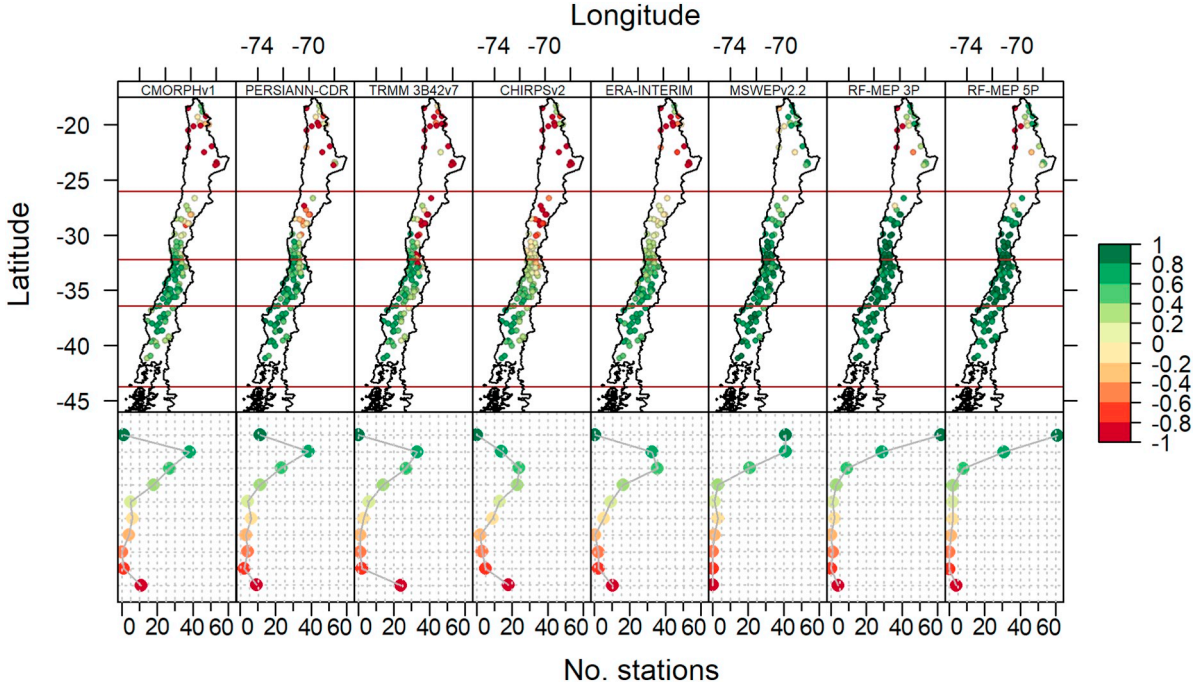

Fig. 7. Spatial distribution of the 3-day KGE' for all *P* products using ground-based measurements. The dotplots in the bottom of the figure show the number of stations from the validation set (111 stations in total) within each KGE' range.

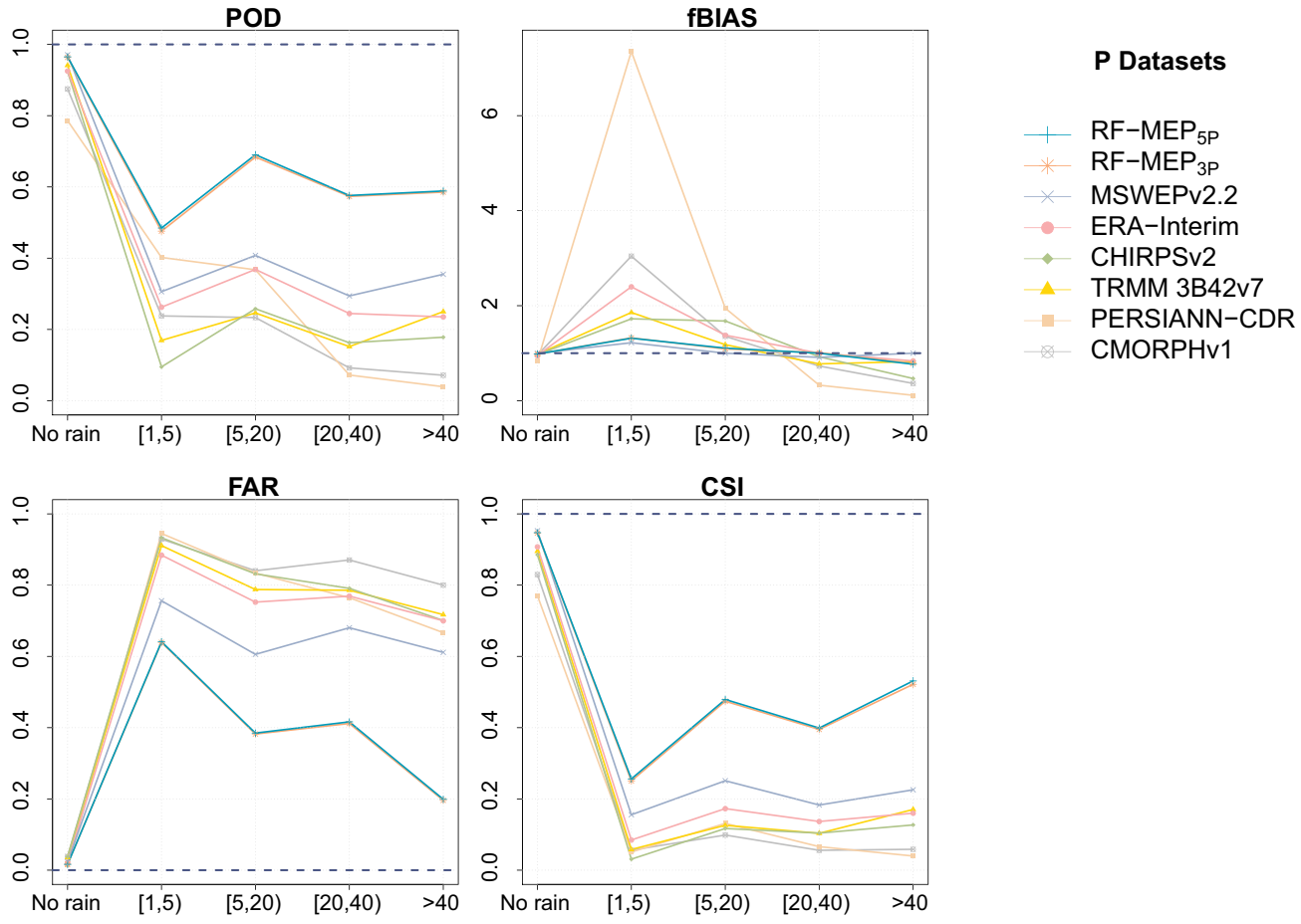

Fig. 8. Median values of the categorical indices of performance at the five *P* intensity classes (in mm d^{-1}) described in Table 3. From left to right and top to bottom: POD, *f*bias, FAR, and CSI. This analysis is biased towards the merged products due to the difference in the reporting times between the rain gauge stations and the *P* products. A square bracket indicates the inclusion of the limit value, while a round bracket indicates its exclusion. The blue line represents the optimal value of each index.

used in their computation at all evaluated temporal scales (see Figs. 4, 5, and Table 4). RF-MEP was able to improve the spatio-temporal representation of *P* (see Figs. 4–8) by combining multiple sources of information. Both merged products showed increased *r*, *β*, and γ values at all temporal scales, which indicates that this method is able to represent the total volume and distribution of *P* by providing a better representation of daily *P* patterns. Comparable improvements in *β* were obtained by Manz et al. (2016) and Yang et al. (2017), although Ma et al. (2018) reported a higher bias in their merged product. Also, the reduction in the dispersion of the KGE' and its components demonstrates that the merged products show good performance over most of the study area. The KGE' has proven to be a useful performance index because of its ability to decompose the performance into *r*, *β*, and γ, which can be used to understand the different sources of mismatches.

The evaluated *P* products showed higher performances at the monthly, seasonal and annual scales in comparison to shorter temporal scales (Fig. 4), similar to the results reported by Jiang et al. (2012) and Zambrano-Bigiarini et al. (2017). This indicates that despite systematic, random, and detection errors present in *P* products at the daily scale, they are still able to represent *P* patterns when aggregated at longer

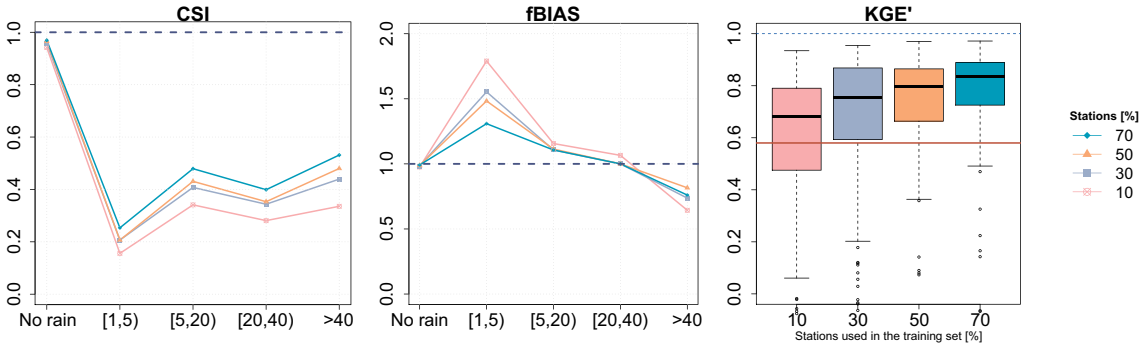

Fig. 9. Performance of the RF-MEP_{5P} using varying percentages of rain gauge stations in the training set. From left to right and top to bottom: the CSI; *f*bias; and the KGE' evaluation of the derived products (the red line indicates the median KGE' of the best performing product used in the computation of RF-MEP_{5P}). The blue line represents the optimal value for each index. A square bracket indicates the inclusion of the limit value, while a round bracket indicates its exclusion.

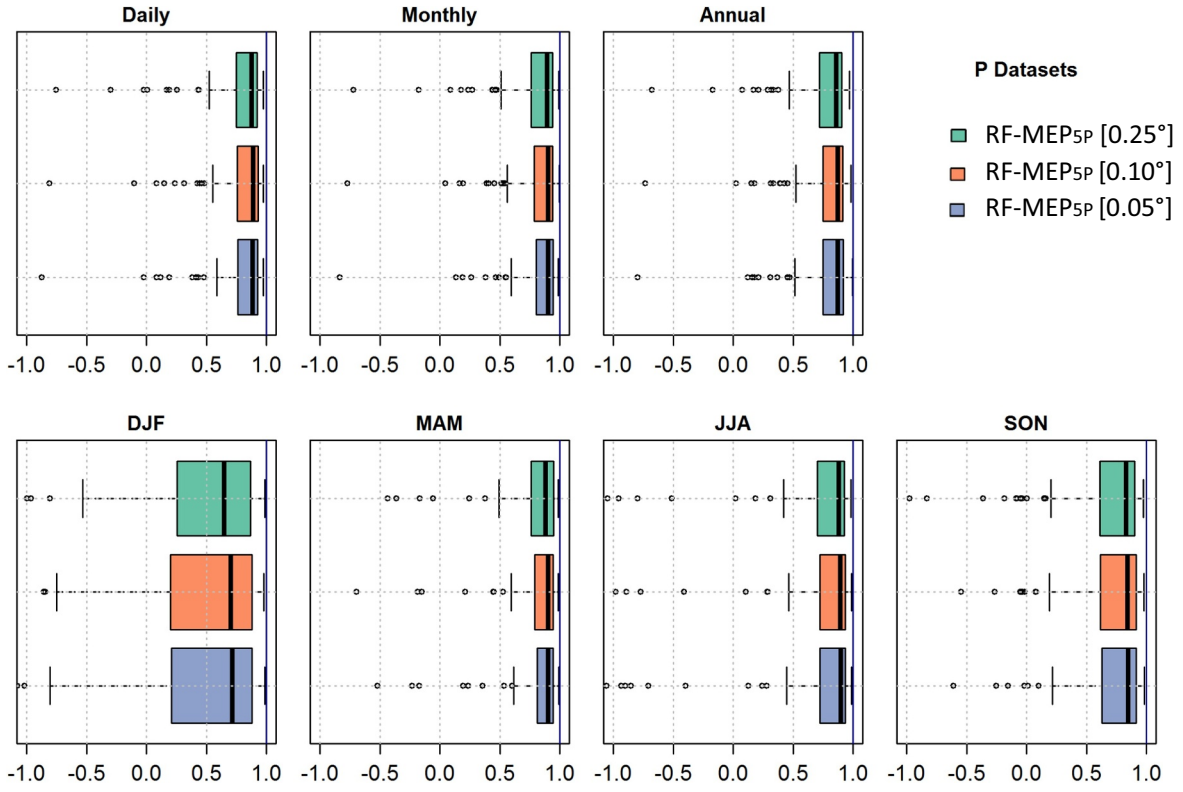

Fig. 10. KGE' values of the RF-MEP$_{5P}$ computed at three spatial resolutions (0.25°, 0.10°, and 0.05°). The solid line represents the median value, the edges of the boxes represent the first and third quartiles, and the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. The blue line shows the optimal value for the KGE'.

temporal scales. On the other hand, Maggioni and Massari (2018) concluded that spatial sampling uncertainties tend to decrease for higher temporal resolutions, which means that the point-to-pixel evaluation tends to be more reliable for increasing accumulation periods.

All products showed the lowest performance in summer (DJF), which is consistent with the results obtained by Rabiei and Haberlandt (2015) and Zambrano-Bigiarini et al. (2017). This could be because: *i*) small-scale convective precipitation events dominate in summer in the Far North region (Prein and Gobiet, 2017); *ii*) in warm months, the evaporation of hydrometeors before they reach the ground leads to overestimation and false alarms (Maggioni and Massari, 2018); and *iii*) passive microwave radiometers overestimate and underestimate *P* during summer and winter, respectively (Tang et al., 2014).

Both merged products presented their lowest performance over the arid Far North region as a consequence of the low performance of all *P* products used as covariates (see Fig. 7). This is in agreement with Manz

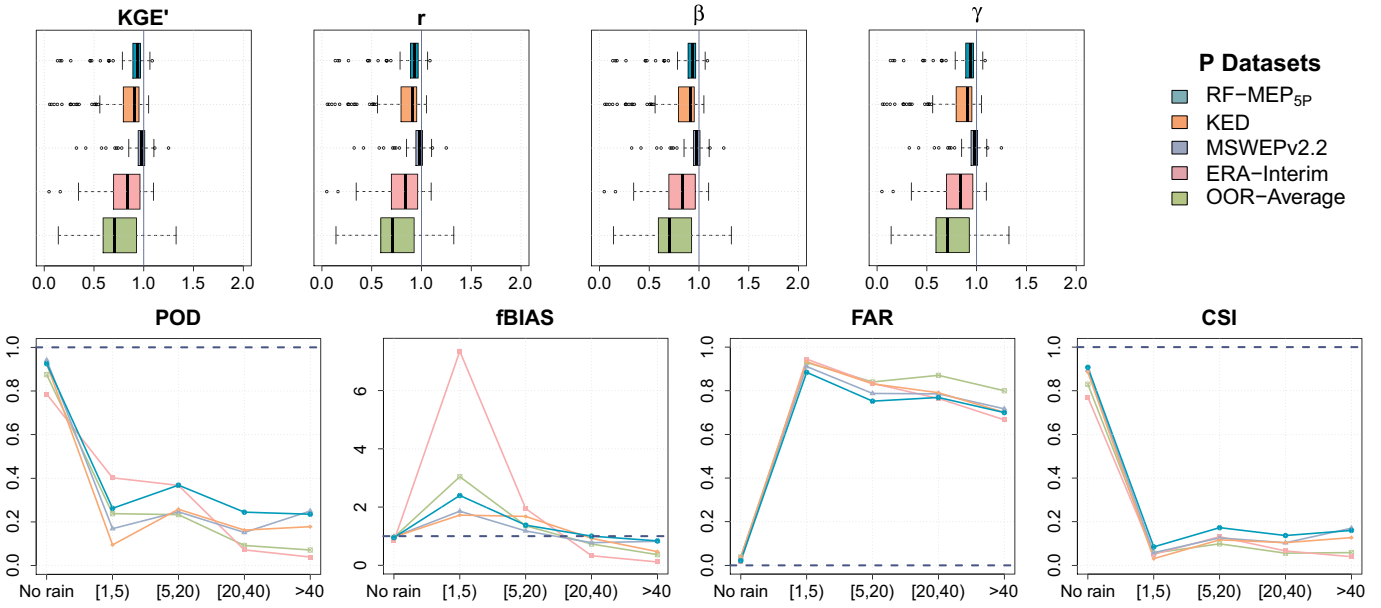

Fig. 11. Performance of different merging procedures using the KGE' and its components as continuous indices and POD, *f*bias, FAR, and CSI as categorical indices at the 3-day temporal scale.

Table 4
Median values of the continuous indices used in the evaluation of *P* products.

P product	KGE'	*r*	*β*	γ
CMORPHv1	0.43	0.67	1.03	0.82
PERSIANN-CDR	0.23	0.62	1.34	0.50
TRMM 3B42v7	0.47	0.69	1.05	0.88
CHIRPSv2	0.48	0.62	1.04	0.71
ERA-Interim	0.58	0.82	1.31	0.84
MSWEPv2.2	0.74	0.89	**1.00**	**0.97**
RF-MEP$_{3P}$	**0.83**	**0.94**	1.03	0.93
RF-MEP$_{5P}$	**0.83**	**0.94**	1.04	0.94

et al. (2016), where the merged products presented high uncertainty and low performances predominantly over regions with low and intermittent *P* regimes. The mismatches of the *P* products are more evident in arid and semi-arid climates because over low *P* regimes, any overestimation or underestimation will have a greater impact on the performance evaluation. Despite this, the RF-MEP$_{5P}$ and RF-MEP$_{3P}$ products were able to adequately represent the *P* patterns of the higher elevations of the Far North, showing that RF-MEP is able to improve the spatio-temporal estimation of *P* through the inclusion of complementary information, even in regions where the selected products exhibit low performance.

Because both merged products were computed using daily gauge data from the national water agencies they represent daily accumulations from 11:00–10:59 UTC, whereas all other selected *P* products represent daily *P* accumulations from 0:00 to 23:59 UTC (~11 h difference; for discussion, see Beck et al., 2019). This time difference must be considered for the evaluation of the *P* products at the daily temporal scale. Among the evaluated *P* products, only MSWEPv2.2 incorporates daily gauge data and applies corrections to account for the reporting times of the rain gauges. Fig. 12 shows the evaluation of the *P* products for 1-day and 3-day periods. Both merged products performed similarly well with a median KGE' of 0.83 because they use the Chilean rain gauges; however, the five *P* products used in their computation performed slightly worse in the 1-day evaluation due to the 11 h difference in the reporting times. The 3-day temporal scale was considered sufficient to render the difference in reporting times negligible.

5.2. Correction of mismatches of the original P products

Our results showed that the blending of multiple *P* estimates, topography-related information, and ground-based measurements, can improve the spatio-temporal characterisation of *P*, which is consistent with the results obtained by Verdin et al. (2016) and Manz et al. (2016). The *r*, *β*, and γ components improved at all temporal scales. The γ of both merged products showed a systematic underestimation (γ ~ 0.9, see Fig. 5) at all temporal scales as a consequence of averaging the predictions of the different trees from the RF model. Despite this, Fig. 5c demonstrates that the γ values of the merged products are higher than those shown by the products used as covariates.

Recently, Alvarez-Garreton et al. (2018) derived runoff coefficients larger than 1, mainly over Central Chile and in the Far-South, with increasing coefficient values towards the Andes. This finding is consistent with those of Beck et al. (2017a), indicating that more water is leaving the catchments than the total amount entering as *P*. This suggests that the *P* products systematically underestimate *P* at high elevations throughout Chile, which may be due to the inability of satellite-based products to accurately estimate *P* over snow and ice-covered surfaces (Beck et al., 2017a). Also, during winter, most Chilean rain gauges located at high elevations are not able to correctly incorporate snow into the *P* measurement, leading to an underestimation of *P*. Therefore, even considering the good performance of the two merged products at different temporal scales, it is likely that the real amount of *P* is underestimated at high elevations due to the absence of ground-based information. To reduce the possible underestimation of *P* over high elevation and snow-driven catchments, the incorporation of rain gauges able to measure both liquid and solid precipitation at high elevations is recommended, along with the use of *P* products that account for solid *P* (such as MSWEPv2.2 and reanalysis products).

The inclusion of different *P* products improved the detection of different *P* intensities at the daily scale, as observed in the improved categorical performance of the merged products compared to that of the covariates (see Fig. 8 and Table 5). The categorical performance of both merged products showed an improved detection of the selected *P* intensities and a reduction in the amount of days that are incorrectly classified. These results, in combination with the improved values of *r* and *β*, show that RF-MEP is capable of correcting *P* events at the daily

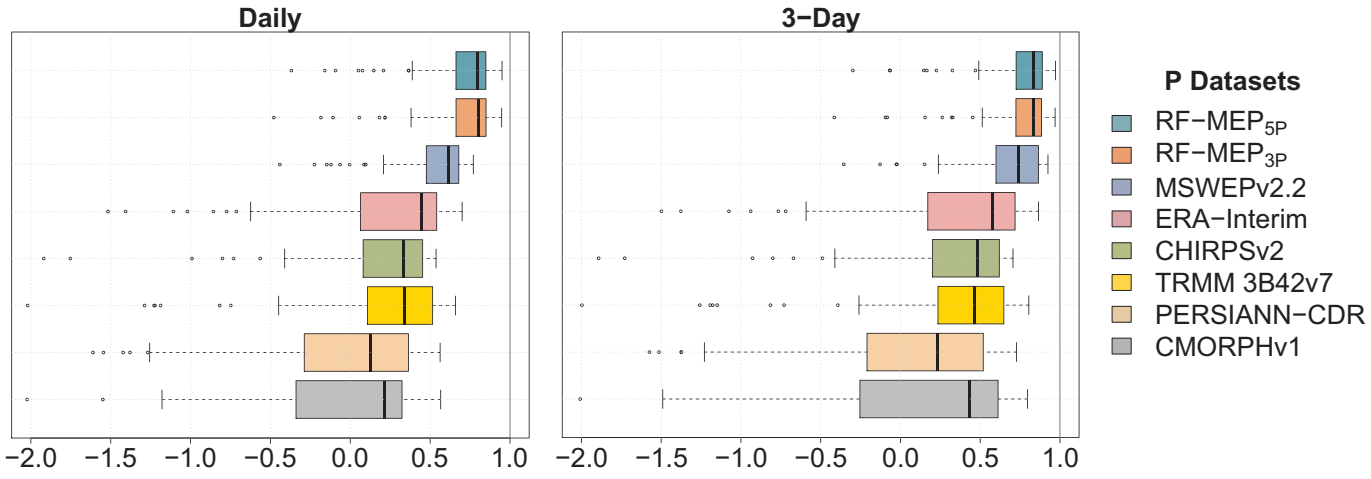

Fig. 12. KGE' values calculated using the ground-based validation dataset at the 1-day time scale (left) and the 3-day time scale (right). The solid line represents the median value, the edges of the boxes represent the first and third quartiles, and the whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the box. The vertical blue line indicates the optimal value for KGE'.

Table 5
Median values of POD, FAR, *f*bias, and CSI for the different *P* intensities (see Table 3) for ERA-Interim, MSWEPv2.2, and RF-MEP$_{5P}$.

Intensity	ERA-Interim				MSWEPv2.2				RF-MEP$_{5P}$			
(mm)	POD	FAR	*f*bias	CSI	POD	FAR	*f*bias	CSI	POD	FAR	*f*bias	CSI
[0, 1)	0.92	0.02	0.95	0.91	0.97	0.02	1.00	0.95	0.96	0.02	0.98	0.95
[1, 5)	0.26	0.88	2.40	0.08	0.30	0.76	1.22	0.15	0.48	0.64	1.31	0.26
[5, 20)	0.37	0.75	1.38	0.17	0.40	0.60	1.00	0.25	0.69	0.38	1.10	0.48
[20, 40)	0.24	0.77	1.00	0.14	0.29	0.68	0.91	0.18	0.58	0.42	1.00	0.40
≥ 40	0.23	0.70	0.83	0.16	0.35	0.61	1.00	0.22	0.59	0.20	0.77	0.53

scale, assigning more accurate *P* amounts to each day, and preserving the total volume of *P* at larger scales; consequently improving the spatial representation of *P* patterns.

The analysis of the *P* products at different intensities is affected by the difference in reporting times between the products and the ground-based measurements (see Fig. 12). All the products used as covariates, with the exception of CHIRPSv2 and TRMM 3B42v7, presented statistically significant differences at the 95% confidence interval between the daily and 3-day values. This issue is unfortunately ignored in the majority of *P* evaluation studies and constitutes a major limitation of most evaluations carried out in time zones far from 0:00 UTC.

Fig. 13 shows the relative difference of mean annual *P* (2000–2016) between each product and the values observed at the rain gauges of the validation set. These values are in agreement with the spatial performance assessment (Fig. 7), where the *P* products presented the lowest performance in the Far North. The blue colours indicate overestimation of the products, while the red colours indicate underestimation. *P* is overestimated in the Far North by CMORPHv1, PERSIANN-CDR, TRMM 3B42v7, CHIRPSv2, and ERA-Interim; and as a consequence, both merged products overestimate *P* over this region (except for the high elevated areas). These results are in agreement with Dinku et al. (2011) and Zambrano-Bigiarini et al. (2017), where the products overestimated *P* over the arid regions of Africa and Chile, respectively. MSWEPv2.2 and the merged products were able to capture the *P* volume over the mountainous area in the Far North, despite the challenges presented by climate variability caused by extreme topography and by a lack of ground-based measurements (Maggioni and Massari, 2018).

The merged products show lower relative difference, i.e. good performance, for almost all stations in the Near South, Central Chile, South, and elevated areas in the Far North. The improved performance of the merged products can be observed in the lower panel of Fig. 13, which highlights that the majority of the *P* products presented relative differences between −0.2 and 0.2 compared to rain gauges. This suggests that RF-MEP is capable of representing the mean annual *P* patterns when applied at daily temporal scale. The overestimation over the Far North is expected because all products used to derive both merged products tend to overestimate *P* over this region.

5.3. Impact of network density, spatial resolution, and limitations

A high number of rain gauge stations in the training set leads to higher performance and higher detection of *P* intensities, as observed in Fig. 9. When we reduced the training sample to 10% (37) of the total

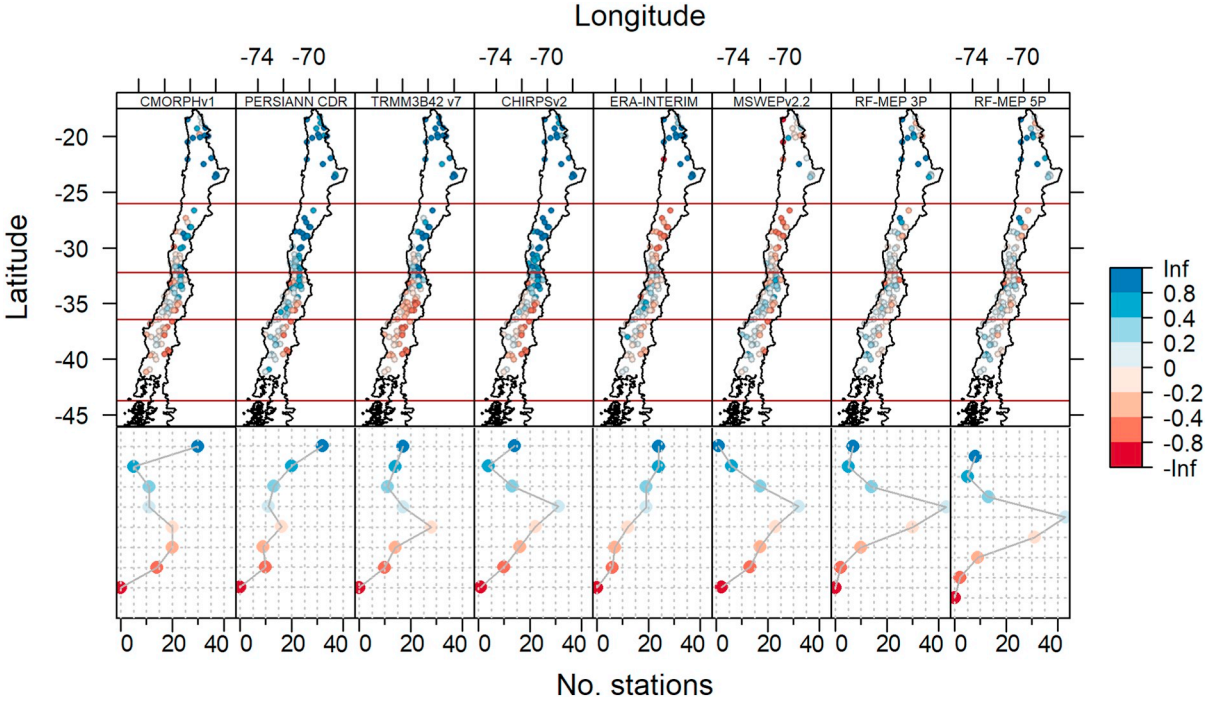

Fig. 13. Mean annual relative difference for the *P* products for 2000–2016. The points with negative values (red colours) are underestimated by the respective product, while the points with positive values (blue colours) are overestimated.

available stations, RF-MEP$_{5P}$ was still able to outperform the products used as covariates, showing the effectiveness of the proposed RF-MEP method.

The products RF-MEP$_{5P}$ and RF-MEP$_{3P}$ performed similarly, as observed in Figs. 4–8. The median values and the interquantile ranges of the KGE', r, β, and γ are similar for both merged products, except over the Far North, where RF-MEP$_{5P}$ shows less dispersion in the KGE' and its components than RF-MEP$_{3P}$, despite the slight decrease in the median performance. This indicates that the inclusion of more P products could reduce the dispersion in areas where the selected products show low performance. The similar performance of RF-MEP$_{5P}$ and RF-MEP$_{3P}$ indicates that the method is able to extract useful information from the P products. Similar results were obtained when RF-MEP$_{3P}$ used ERA-Interim, CMORPHv1, and TRMM 3B42v7 instead of ERA-Interim, CHIRPSv2, and PERSIANN-CDR (please see Figs. S2 and S3 from the supplementary material), demonstrating that RF-MEP is a robust merging method. Although the P products must be resampled to the same spatial resolution to generate the merged product, the effect of including P products generated at different spatial resolutions is negligible (see Fig. 10).

RF-MEP$_{5P}$ includes CMORPHv1 and TRMM 3B42v7, which reduces the potential temporal coverage by 15 years (RF-MEP$_{3P}$ can be generated from 1983 onwards, while RF-MEP$_{5P}$ can only be generated from 1998). Therefore, based on the similar strong performances of both merged products (see Section 5.1), we prefer RF-MEP$_{3P}$ for the Chilean case study, as the benefits of including CMORPHv1 and TRMM 3B42v7 to generate RF-MEP$_{5P}$ are outweighed by the loss of 15 years of record.

Although RF-MEP was only applied over Chile, we are confident that this method could be successfully applied over other areas, due to its outstanding performance in a region with notable heterogeneity in topography and climate, and because it was able to improve the spatio-temporal characterisation of P even when the training set was largely reduced. However, some limitations apply to this method: *i*) since ground-based data are necessary, it would be difficult to apply the proposed method globally and in near-real time; *ii*) it can be computationally intensive when applied to large areas; and *iii*) it has problems predicting values that are completely out from the training range.

6. Conclusion

Satellite and reanalysis-based P estimates provide an unprecedented opportunity for numerous hydrological, meteorological, and other environmental applications. Despite the continuous improvements of P products, different types of mismatches still exist in most of them. Here we present RF-MEP, a novel method capable of deriving improved P estimates by merging information from (near-)global and publicly available P products, rain gauge stations, and topography-related data. Two merged products (RF-MEP$_{3P}$ and RF-MEP$_{5P}$) obtained with the proposed method showed improved r, β, and γ values at all temporal scales compared to all the individual P products used as covariates. Furthermore, both merged datasets exhibited improved POD, FAR, $f$bias, and CSI for different P intensities. Finally, both merged products performed better than the benchmark dataset MSWEPv2.2, except during summer (DJF). The key findings of the application of this method to the Chilean case study are as follows:

(a) RF-MEP can be applied at different temporal scales (e.g., daily, monthly, or annually) to obtain an improved spatio-temporal representation of P patterns.
(b) The different P products used in this study performed better at longer timescales than at short timescales, while both merged products performed well at all timescales.
(c) RF-MEP$_{3P}$ and RF-MEP$_{5P}$ outperformed all the evaluated P products at the 3-day, monthly, annual, MAM, JJA, and SON temporal scales. However, the benchmark MSWEPv2.2 outperformed the merged products during summer (DJF).
(d) RF-MEP$_{3P}$ (which uses CHIRPSv2, PERSIANN-CDR, and ERA-Interim) showed a similar performance to RF-MEP$_{5P}$ (which also included CMORPHv1 and TRMM 3B42v7). Therefore, including CMORPHv1 and TRMM 3B42v7 as covariates in the merging procedure only led to a minor increase in the overall performance of the final merged product. Consequently, for the Chilean case study, it is preferable to use RF-MEP$_{3P}$ and gain 15 years of data (1983 as the starting date instead of 1998).
(e) The performance of RF-MEP increases when more rain gauge stations are used to train the model; however, it is still able to improve P characteristics even with relatively few stations in the training set.
(f) RF-MEP showed better performance than the results obtained using Kriging with external drift and one-outlier-removed arithmetic mean.
(g) The difference in reporting times between the P products and the ground-based measurements must be taken into account when assessing the performance of P products at the daily temporal scale so that their performance is not underestimated. This issue constitutes a major limitation of most P evaluation studies carried out far from 0:00 UTC.
(h) The KGE' proved to be a versatile performance index because of its ability to decompose the performance of the P products into r, β, and γ. Therefore, the KGE' helps us understand the sources of mismatches between the P products and ground-based observations. In addition, the use of categorical indices provides crucial information about the performance of these P datasets for capturing different P intensities.

RF-MEP was developed to improve the characterisation of the spatio-temporal variability of P by merging multiple P products, topography-related datasets, and ground-based information. The P products used in this study are publicly available and have a (quasi-)global spatial coverage. This method was validated over Chile, a country which exhibits notable heterogeneity in topography, climate, and land cover. For this reason, we are confident that RF-MEP can be successfully applied in different regions and catchments worldwide, and could also be used to improve other climatological variables when ground-based data are available.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors thank the Centers for Natural Resources and Development (CNRD) Ph.D. program for their financial support to the main author. The daily RF-MEP$_{3P}$ will be made available upon request for 2000–2016 at a spatial resolution of 0.05°. In addition, Dr. Zambrano-Bigiarini thanks Conicyt-Fondecyt 11150861 for the financial support from 2016 to 2018 and the Center for Climate and Resilience Research (CR2, Conicyt-FONDAP 15110009) for providing the rain gauge data. We particularly thank Tim R. McVicar and the three anonymous reviewers, whose constructive comments helped to improve the quality of the final manuscript. We would also like to thank the RSE editorial team for their support.

Appendix A. Literature review table

Table A1
Main steps in the methodology of different studies that have applied merging algorithms to improve the spatio-temporal characterisation of *P* at different temporal scales.

Study	Merging method(s)	Spatio-temporal resolution of the merged product(s)	Description of the approach
Li and Shao (2010)	Nonparametric kernel smoothing	Daily (0.25°)	1. Calculation of residual values; 2. Background error estimation using a kernel smoothing method (double smoothing); 3. Removal of the estimated error from the background field
Rozante et al. (2010)	Barnes objective analysis method	Daily (0.25°)	1. Only the rain gauge observations are considered over the 5 by 5 square of cells centred around every grid-cell with a rain gauge station; 2. Interpolation using the Barnes objective analysis method for the remaining grid-cells
Xie and Xiong (2011)	Optimal Interpolation	Daily (0.25°)	1. Bias correction through a probability density function matching of satellite and rain gauge data; 2. Optimal interpolation
Gebregiorgis and Hossain (2011)	Linear weights based on hydrologic model predictability	Daily (0.125°)	1. Calculation of the mean squared error (MSE) of soil moisture and runoff using each *P* product to force a distributed hydrological model; 2. Inversion of MSEs to be used as weights; 3. Merging of the *P* products using linear weighting
Woldemeskel et al. (2013)	Linearised weighting procedure	Monthly (0.05)	1. *P* interpolation using thin plate smoothing splines (TPSS) with standardised rain gauge data followed by a back-transformation; 2. Merging using a linearised weighting procedure
Shen et al. (2014)	Arithmetic mean and inverse-error-square weighting methods	Daily (0.25°)	Three methods: M1. Arithmetic mean; M2. Inverse-error-square weighting; M3. One-outlier removed arithmetic mean (i.e., one product removed)
Nie et al. (2015)	Optimal interpolation	Daily (0.25°)	1. Bias correction through a cumulative distribution function matching procedure; 2. Quantification of background and observation errors; 3. Application of the optimal interpolation technique
Fu et al. (2016)	Bayesian model averaging	Annual mean (0.1°)	1. Non-linear spatial interpolation of *P* products; 2. Merging using the Bayesian model averaging technique
Manz et al. (2016)	Linear modelling, residual IDW, and Kriging-based methods	Monthly mean (5 km ~ 0.05°)	Five methods: M1. Linear Modeling; M2. Residual IDW; M3. Ordinary Kriging (only gauge-based); M4. Residual ordinary Kriging; M5. Kriging with external drift
Verdin et al. (2016)	Ordinary Kriging and k-nearest neighbour local polynomials	Monthly (0.05°)	Two methods: M1. Ordinary Kriging; M2. A local regression is fitted considering data from within a small neighbourhood, and the weighted least squares are used to fit the local polynomials
Shi et al. (2017)	Merging weights based on the effective influence radius of rain gauges	Hourly (1 km)	1. Selection of the *P* product; 2. Downscaling of the *P* product using a DEM; 3. Determination of weighted differences between the downscaled product and rain gauge data; 4. Merging the downscaled product and the weighted differences considering the number of gauges in the effective influence radius
Yang et al. (2017)	Inverse-root-mean-square-error weighting	Daily (0.04°)	1. Bias correction of the *P* product using a quantile mapping technique and a Gaussian weighting interpolation scheme; 2. Interpolation of rain gauge data using a Gaussian weighting function; 3. Data merging using inverse-mean-square-error weighting
Ma et al. (2018)	Bayesian Model Averaging	Daily (0.25°)	1. A BMA scheme is used to adjust the PDF of the satellite estimates with the expectation-maximisation method used for each member for each day at the gauge locations; 2. Interpolation using OK
Beck et al. (2019)	Weighted averaging with CDF matching	3-hourly (0.10°)	1. Gauge data quality control; 2. Inferring gauge reporting times; 3. Rainfall estimation using thermal infrared imagery; 4. Gauge-based assessment of satellite and reanalysis *P* datasets; 5. Global maps of weights and wet-day biases; 6. Determination of long-term mean *P*; 7. *P* frequency correction and dataset harmonisation; 8. Reference *P* distributions; 9. Merging of satellite and reanalysis *P* datasets; 10. Gauge correction scheme

Appendix B. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rse.2019.111606. These data include the Google maps of the most important areas described in this article.

References

Adhikary, S.K., Yilmaz, A.G., Muttil, N., 2015. Optimal design of rain gauge network in the Middle Yarra River catchment, Australia. Hydrol. Process. 29, 2582–2599.

Alvarez-Garreton, C., Mendoza, P.A., Boisier, J.P., Addor, N., Galleguillos, M., Zambrano-Bigiarini, M., Lara, A., Cortes, G., Garreaud, R., McPhee, J., et al., 2018. The CAMELS-CL dataset: catchment attributes and meteorology for large sample studies-Chile dataset. Hydrol. Earth Syst. Sci. 22, 5817–5846.

Ashouri, H., Hsu, K.-L., Sorooshian, S., Braithwaite, D.K., Knapp, K.R., Cecil, L.D., Nelson, B.R., Prat, O.P., 2015. PERSIANN-CDR: daily precipitation climate data record from multisatellite observations for hydrological and climate studies. Bull. Am. Meteorol. Soc. 96, 69–83.

Baez-Villanueva, O.M., Zambrano-Bigiarini, M., Ribbe, L., Nauditt, A., Giraldo-Osorio, J.D., Thinh, N.X., 2018. Temporal and spatial evaluation of satellite rainfall estimates over different regions in Latin-America. Atmos. Res. 213, 34–50.

Beck, H.E., van Dijk, A.I., De Roo, A., Miralles, D.G., McVicar, T.R., Schellekens, J., Bruijnzeel, L.A., 2016. Global-scale regionalization of hydrologic model parameters. Water Resour. Res. 52, 3599–3622.

Beck, H.E., van Dijk, A.I.J.M., Levizzani, V., Schellekens, J., Miralles, D.G., Martens, B., de Roo, A., 2017a. MSWEP: 3-hourly 0.25° global gridded precipitation (1979-2015) by merging gauge, satellite, and reanalysis data. Hydrol. Earth Syst. Sci. 21, 589–615.

Beck, H.E., Vergopolan, N., Pan, M., Levizzani, V., van Dijk, A.I.J.M., Weedon, G.P., Brocca, L., Pappenberger, F., Huffman, G.J., Wood, E.F., 2017b. Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. Hydrol. Earth Syst. Sci. 21, 6201–6217.

Beck, H.E., Wood, E.F., Pan, M., Fisher, C.K., Miralles, D.G., van Dijk, A.I., McVicar, T.R., Adler, R.F., 2019. MSWEP V2 global 3-hourly 0.1∘ precipitation: methodology and quantitative assessment. Bull. Am. Meteorol. Soc. https://doi.org/10.1175/BAMS-D-17-0138.1.

Beck, H.E., Zimmermann, N.E., McVicar, T.R., Vergopolan, N., Berg, A., Wood, E.F., 2018. Present and future Köppen-Geiger climate classification maps at 1-km resolution. Sci. Data 5, 180–214.

Behrens, T., Schmidt, K., Viscarra Rossel, R.A., Gries, P., Scholten, T., MacMillan, R.A., 2018. Spatial modelling with Euclidean distance fields and machine learning. Eur. J. Soil Sci.

Bergeron, T., 1960. Preliminary results of project pluvius. Publ. Comm. Land Erosion 53, 226–237.

Biau, G., Scornet, E., 2016. A random forest guided tour. TEST 25, 197.
Borga, M., Vizzaccaro, A., 1997. On the interpolation of hydrologic variables: formal equivalence of multiquadratic surface fitting and kriging. J. Hydrol. 195, 160–171.
Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32.
Chen, M., Shi, W., Xie, P., Silva, V.B., Kousky, V.E., Wayne Higgins, R., Janowiak, J.E., 2008. Assessing objective techniques for gauge-based analyses of global daily precipitation. J. Geophys. Res.-Atmos. 113.
Chen, Y., Xia, J., Liang, S., Feng, J., Fisher, J.B., Li, X., Li, X., Liu, S., Ma, Z., Miyata, A., et al., 2014. Comparison of satellite-based evapotranspiration models over terrestrial ecosystems in china. Remote Sens. Environ. 140, 279–293.
Dee, D.P., Uppala, S.M., Simmons, A.J., Berrisford, P., Poli, P., Kobayashi, S., Andrae, U., Balmaseda, M.A., Balsamo, G., Bauer, P., Bechtold, P., Beljaars, A.C.M., van de Berg, L., Bidlot, J., Bormann, N., Delsol, C., Dragani, R., Fuentes, M., Geer, A.J., Haimberger, L., Healy, S.B., Hersbach, H., Hólm, E.V., Isaksen, L., Kållberg, P., Köhler, M., Matricardi, M., McNally, A.P., Monge-Sanz, B.M., Morcrette, J.-J., Park, B.K., 2011. The ERA-Interim reanalysis: configuration and performance of the data assimilation system. Q. J. R. Meteorol. Soc. 137, 553–597.
Derin, Y., Yilmaz, K.K., 2014. Evaluation of multiple satellite-based precipitation products over complex topography. Journal of Hydrometeorology 15, 1498–1516. https://doi.org/10.1175/JHM-D-13-0191.1.
Díaz-Uriarte, R., Alvarez de Andrés, S., 2006. Gene selection and classification of microarray data using random forest. BMC Bioinf. 7, 3.
Dinku, T., Ceccato, P., Connor, S.J., 2011. Challenges of satellite rainfall estimation over mountainous and arid parts of East Africa. Int. J. Remote Sens. 32, 5965–5979.
Dinku, T., Ceccato, P., Grover-Kopec, E., Lemma, M., Connor, S., Ropelewski, C., 2007. Validation of satellite rainfall products over East Africa's complex topography. Int. J. Remote Sens. 28, 1503–1526.
Fu, Y., Xia, J., Yuan, W., Xu, B., Wu, X., Chen, Y., Zhang, H., 2016. Assessment of multiple precipitation products over major river basins of China. Theor. Appl. Climatol. 123, 11–22.
Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., Husak, G., Rowland, J., Harrison, L., Hoell, A., Michaelsen, J., 2015. The climate hazards infrared precipitation with stations-a new environmental record for monitoring extremes. Sci. Data 2, 150066.
Gallant, J., Read, A., Dowling, T., 2012. Removal of tree offsets from SRTM and other digital surface models. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. 39, 275–280.
Gao, Y.C., Liu, M.F., 2013. Evaluation of high-resolution satellite precipitation products using rain gauge observations over the Tibetan Plateau. Hydrol. Earth Syst. Sci. 17, 837–849.
Garcia, M., Peters-Lidard, C.D., Goodrich, D.C., 2008. Spatial interpolation of precipitation in a dense gauge network for monsoon storm events in the southwestern united states. Water Resour. Res. 44.
Gebregiorgis, A., Hossain, F., 2011. How much can a priori hydrologic model predictability help in optimal merging of satellite precipitation products? J. Hydrometeorol. 12, 1287–1298.
Goudenhoofdt, E., Delobbe, L., 2009. Evaluation of radar-gauge merging methods for quantitative precipitation estimates. Hydrol. Earth Syst. Sci. 13, 195–203.
Grimes, D.I.F., Pardo-Igúzquiza, E., Bonifacio, R., 1999. Optimal areal rainfall estimation using raingauges and satellite data. J. Hydrol. 222, 93–108.
Gupta, H.V., Kling, H., Yilmaz, K.K., Martinez, G.F., 2009, Oct, Oct. Decomposition of the mean squared error and NSE performance criteria: implications for improving hydrological modelling. J. Hydrol. 377, 80–91.
Hengl, T., 2019. GSIF: Global Soil Information Facilities. R package version 0.5-5. .
Hengl, T., Nussbaum, M., Wright, M.N., Heuvelink, G.B., Gräler, B., 2018. Random Forest as a generic framework for predictive modeling of spatial and spatio-temporal variables. PeerJ 6, e5518.
Hijmans, R.J., 2018. raster: Geographic Data Analysis and Modeling. R package version 2.8-4. .
Hirpa, F.A., Gebremichael, M., Hopson, T., 2010. Evaluation of high-resolution satellite precipitation products over very complex terrain in Ethiopia. J. Appl. Meteorol. Climatol. 49, 1044–1051.
Huffman, G., Adler, R., Bolvin, D., Nelkin, E., 2010. The TRMM Multi-satellite Precipitation Analysis (TMPA). Chapter 1 in Satellite Rainfall Applications for Surface Hydrology, f. hossain and m. gebremichael, eds.
Huffman, G.J., Bolvin, D.T., Nelkin, E.J., Wolff, D.B., Adler, R.F., Gu, G., Hong, Y., Bowman, K.P., Stocker, E.F., 2007. The TRMM Multisatellite Precipitation Analysis (TMPA): quasi-global, multiyear, combined-sensor precipitation estimates at fine scales. J. Hydrometeorol. 8, 38.
Hutchinson, M.F., 1995. Interpolating mean rainfall using thin plate smoothing splines. Int. J. Geogr. Inf. Syst. 9, 385–403.
Jaagus, J., Briede, A., Rimkus, E., Remm, K., 2010. Precipitation pattern in the baltic countries under the influence of large-scale atmospheric circulation and local landscape factors. Int. J. Climatol. 30, 705–720.
Jarvis, A., Reuter, H., Nelson, A., Guevara, E., 2008. Hole-filled SRTM for the globe version 3, from the CGIAR-CSI SRTM 90m database.
Jiang, S., Ren, L., Hong, Y., Yong, B., Yang, X., Yuan, F., Ma, M., 2012. Comprehensive evaluation of multi-satellite precipitation products with a dense rain gauge network and optimally merging their simulated hydrological flows using the Bayesian model averaging method. J. Hydrol. 452, 213–225.
Joyce, R.J., Janowiak, J.E., Arkin, P.A., Xie, P., 2004. CMORPH: a method that produces global precipitation estimates from passive microwave and infrared data at high spatial and temporal resolution. J. Hydrol. 5, 487–503.
Kling, H., Fuchs, M., Paulin, M., 2012. Runoff conditions in the upper Danube basin under an ensemble of climate change scenarios. J. Hydrol. 424, 264–277.
Li, M., Shao, Q., 2010. An improved statistical approach to merge satellite rainfall estimates and rain gauge data. J. Hydrol. 385, 51–64.
Li, Z., Yang, D., Hong, Y., 2013. Multi-scale evaluation of high-resolution multi-sensor blended global precipitation products over the yangtze river. J. Hydrol. 500, 157–169.
Liaw, A., Wiener, M., 2002. Classification and regression by random forest. R News 2, 18–22.
Lievens, H., Tomer, S.K., Al Bitar, A., De Lannoy, G.J., Drusch, M., Dumedah, G., Franssen, H.-J.H., Kerr, Y.H., Martens, B., Pan, M., et al., 2015. SMOS soil moisture assimilation for improved hydrologic simulation in the Murray Darling Basin, Australia. Remote Sens. Environ. 168, 146–162.
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. Hydrol. Earth Syst. Sci. 15, 2259–2274.
Ma, Y., Hong, Y., Chen, Y., Yang, Y., Tang, G., Yao, Y., Long, D., Li, C., Han, Z., Liu, R., 2018. Performance of optimally merged multisatellite precipitation products using the dynamic Bayesian model averaging scheme over the Tibetan Plateau. J. Geophys. Res.-Atmos. 123, 814–834.
Maggioni, V., Massari, C., 2018. On the performance of satellite precipitation products in riverine flood modeling: a review. J. Hydrol. 558, 214–224.
Manz, B., Buytaert, W., Zulkafli, Z., Lavado, W., Willems, B., Robles, L.A., Rodríguez-Sánchez, J.-P., 2016. High-resolution satellite-gauge merged precipitation climatologies of the Tropical Andes. J. Geophys. Res.-Atmos. 121, 1190–1207.
McKenney, D.W., Pedlar, J.H., Papadopol, P., Hutchinson, M.F., 2006. The development of 1901–2000 historical monthly climate models for Canada and the United States. Agric. For. Meteorol. 138, 69–81.
McVicar, T.R., Körner, C., 2013. On the use of elevation, altitude, and height in the ecological and climatological literature. Oecologia 171, 335–337.
McVicar, T.R., Van Niel, T.G., Li, L., Hutchinson, M.F., Mu, X., Liu, Z., 2007. Spatially distributing monthly reference evapotranspiration and pan evaporation considering topographic influences. J. Hydrol. 338, 196–220.
Montecinos, A., Aceituno, P., 2003. Seasonality of the ENSO-related rainfall variability in Central Chile and associated circulation anomalies. J. Clim. 16, 281–296.
Nie, S., Luo, Y., Wu, T., Shi, X., Wang, Z., 2015. A merging scheme for constructing daily precipitation analyses based on objective bias-correction and error estimation techniques. J. Geophys. Res.-Atmos. 120, 8671–8692.
Oliver, M., Webster, R., 2014. A tutorial guide to geostatistics: computing and modelling variograms and kriging. Catena 113, 56–69.
Prasad, A.M., Iverson, L.R., Liaw, A., 2006. Newer classification and regression tree techniques: bagging and Random Forests for ecological prediction. Ecosystems 9, 181.
Prein, A.F., Gobiet, A., 2017. Impacts of uncertainties in european gridded precipitation observations on regional climate analysis. Int. J. Climatol. 37, 305–327.
R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria.
Rabiei, E., Haberlandt, U., 2015. Applying bias correction for merging rain gauge and radar data. J. Hydrol. 522, 544–557.
Robertson, A.W., Baethgen, W., Block, P., Lall, U., Sankarasubramanian, A., de Souza Filho, F.d.A., Verbist, K.M., 2014. Climate risk management for water in semi-arid regions. Earth Perspect. 1, 12.
Roy, M.-H., Larocque, D., 2012. Robustness of random forests for regression. Journal of Nonparametric Statistics 24, 993–1006.
Rozante, J.R., Moreira, D.S., de Goncalves, L.G.G., Vila, D.A., 2010. Combining TRMM and surface observations of precipitation: technique and validation over south america. Weather Forecast. 25, 885–894.
Schneider, U., Fuchs, T., Meyer-Christoffer, A., Rudolf, B., 2008. Global precipitation analysis products of the GPCC. 112 Global Precipitation Climatology Centre (GPCC), DWD: Internet Publication.
Seo, D.-J., Krajewski, W.F., Bowles, D.S., 1990. Stochastic interpolation of rainfall data from rain gages and radar using cokriging: 1. Design of experiments. Water Resour. Res. 26, 469–477.
Shen, Y., Xiong, A., Hong, Y., Yu, J., Pan, Y., Chen, Z., Saharia, M., 2014. Uncertainty analysis of five satellite-based precipitation products and evaluation of three optimally merged multi-algorithm products over the Tibetan Plateau. Int. J. Remote Sens. 35, 6843–6858.
Shi, H., Chen, J., Li, T., Wang, G., 2017. A new method for estimation of spatially distributed rainfall through merging satellite observations, rain gauge records, and terrain digital elevation model data. J. Hydro Environ. Res.
Sinclair, S., Pegram, G., 2005. Combining radar and rain gauge rainfall estimates using conditional merging. Atmos. Sci. Lett. 6, 19–22.
Sorooshian, S., Hsu, K.-L., Gao, X., Gupta, H.V., Imam, B., Braithwaite, D., 2000. Evaluation of PERSIANN system satellite-based estimates of tropical rainfall. Bull. Am. Meteorol. Soc. 81, 2035–2046.
Tang, L., Tian, Y., Lin, X., 2014. Validation of precipitation retrievals over land from satellite-based passive microwave sensors. J. Geophys. Res.-Atmos. 119, 4546–4567.
Thiemig, V., Rojas, R., Zambrano-Bigiarini, M., De Roo, A., 2013. Hydrological evaluation of satellite-based rainfall estimates over the volta and Baro-Akobo Basin. J. Hydrol. 499, 324–338.
Thiemig, V., Rojas, R., Zambrano-Bigiarini, M., Levizzani, V., De Roo, A., 2012. Validation of satellite-based precipitation products over sparsely gauged African river basins. J. Hydrometeorol. 13, 1760–1783.
Valdés-Pineda, R., Pizarro, R., García-Chevesich, P., Valdés, J.B., Olivares, C., Vera, M., Balocchi, F., Pérez, F., Vallejos, C., Fuentes, R., Abarza, A., Helwig, B., 2014. Water governance in Chile: availability, management and climate change. J. Hydrol. 519, 2538–2567.
Verbist, K., Robertson, A.W., Cornelis, W.M., Gabriels, D., 2010. Seasonal predictability of daily rainfall characteristics in central northern Chile for dry-land management. J.

Appl. Meteorol. Climatol. 49, 1938–1955.
Verdin, A., Funk, C., Rajagopalan, B., Kleiber, W., 2016. Kriging and local polynomial methods for blending satellite-derived and gauge precipitation estimates to support hydrologic early warning systems. IEEE Trans. Geosci. Remote Sens. 54, 2552–2562.
Villarini, G., Krajewski, W.F., 2008. Empirically-based modeling of spatial sampling uncertainties associated with rainfall measurements by rain gauges. Adv. Water Resour. 31, 1015–1023.
Villarini, G., Mandapaka, P.V., Krajewski, W.F., Moore, R.J., 2008. Rainfall and sampling uncertainties: a rain gauge perspective. J. Geophys. Res.-Atmos. 113.
Wang, C., Tang, G., Han, Z., Guo, X., Hong, Y., 2018. Global intercomparison and regional evaluation of GPM imerg version-03, version-04 and its latest version-05 precipitation products: similarity, difference and improvements. J. Hydrol. 564, 342–356.
Weiss, L.L., Wilson, W.T., 1953. Evaluation of significance of slope changes in double-mass curves. Trans. Am. Geophys. Union 34, 893–896.
Woldemeskel, F.M., Sivakumar, B., Sharma, A., 2013. Merging gauge and satellite rainfall with specification of associated uncertainty across Australia. J. Hydrol. 499, 167–176.
Xie, P., Joyce, R., Wu, S., Yoo, S.-H., Yarosh, Y., Sun, F., Lin, R., 2017. Reprocessed, bias-corrected CMORPH global high-resolution precipitation estimates from 1998. J. Hydrometeorol. 18, 1617–1641.
Xie, P., Xiong, A.-Y., 2011. A conceptual model for constructing high-resolution gauge-satellite merged precipitation analyses. J. Geophys. Res.-Atmos. 116, D21106.
Yang, Z., Hsu, K., Sorooshian, S., Xu, X., Braithwaite, D., Zhang, Y., Verbist, K.M.J., 2017. Merging high-resolution satellite-based precipitation fields and point-scale rain gauge measurements - a case study in Chile. J. Geophys. Res.-Atmos. 122, 5267–5284.
Zambrano-Bigiarini, M., 2017a. hydroGOF: goodness-of-fit functions for comparison of simulated and observed hydrological time series. R package version 0.3-10. https://doi.org/10.5281/zenodo.840087. http://hzambran.github.io/hydroGOF/.
Zambrano-Bigiarini, M., 2017b. hydroTSM: Time Series Management, Analysis and Interpolation for Hydrological Modelling. R package version 0.5-1. https://doi.org/10.5281/zenodo.83964. https://github.com/hzambran/hydroTSM.
Zambrano-Bigiarini, M., 2018. Temporal and spatial evaluation of long-term satellite-based precipitation products across the complex topographical and climatic gradients of chile. In: Remote Sensing and Modeling of the Atmosphere, Oceans, and Interactions VII International Society for Optics and Photonics, pp. 1078202. https://doi.org/10.1117/12.2513645.
Zambrano-Bigiarini, M., Nauditt, A., Birkel, C., Verbist, K., Ribbe, L., 2017. Temporal and spatial evaluation of satellite-based rainfall estimates across the complex topographical and climatic gradients of Chile. Hydrol. Earth Syst. Sci. 21, 1295–1320. https://doi.org/10.5194/hess-21-1295-2017.
Zhao, Y., Feng, D., Yu, L., Wang, X., Chen, Y., Bai, Y., Hernández, H.J., Galleguillos, M., Estades, C., Biging, G.S., Radke, J.D., Gong, P., 2016. Detailed dynamic land cover mapping of Chile: accuracy improvement by integrating multi-temporal data. Remote Sens. Environ. 183, 170–185.