MethodsX 10 (2023) 102071

Contents lists available at ScienceDirect

MethodsX

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

Method Article

Validation of 3-day rainfall forecast at the regional scale

Raquel Melo [a,b], Sandra Oliveira [a,c,*], Ricardo Tomé [d], Alexandre M. Ramos [d,e], Sérgio C. Oliveira [a,c]

[a] Centro de Estudos Geográficos, Instituto de Geografia e Ordenamento do Território, Universidade de Lisboa, Lisbon 1600-276, Portugal
[b] Instituto de Ciências da Terra, Universidade de Évora, Evora 7000-671, Portugal
[c] Associate Laboratory, TERRA, Portugal
[d] Instituto Dom Luiz, Faculdade de Ciências da Universidade de Lisboa, Lisbon 1749-016, Portugal
[e] Institute of Meteorology and Climate Research, Karlsruhe Institute of Technology, Karlsruhe, Germany

ARTICLE INFO

Method name:
Validation of 3-day rainfall forecast at the regional scale

Keywords:
Rainfall forecast
Automatic meteorological stations
Forecast validation
R software

ABSTRACT

Nearly half of the natural disasters in the world are due to hydro-geomorphological hazards. Therefore, rainfall forecast is a key parameter for the implementation of landslides and flash-floods early warning systems. In this work we developed a routine in R software that enables the validation of a 3-day rainfall forecast by comparison with the daily rainfall data recorded in 101 automatic meteorological stations available in mainland Portugal. The routine integrates the pre-processing of base data, the matching between the 3-day rainfall forecast and the daily rainfall registered in the automatic meteorological stations based on sequence of days, the estimation of the difference between the forecasted and the real rainfall values and the computation of error measures, such as the bias, the mean absolute error, the mean absolute percentage error and the root mean square error. The results from the error measures, estimated for the 101 automatic meteorological stations, are then exported to an excel file. The routine is implemented for mainland Portugal and tested using data from February 2015, however, the spatial and temporal data can be easily updated for other regions.

- A routine to validate the rainfall forecast at the regional scale using R programming language is implemented.
- The automated routine can be easily updated and adapted with different spatial and temporal scales.

Specifications Table

Subject area:	Earth and Planetary Sciences
More specific subject area:	*Spatial and time-series data analysis*
Name of your method:	*Validation of 3-day rainfall forecast at the regional scale*
Name and reference of original method:	*Not applicable*
Resource availability:	R script available as supplementary material

* Corresponding author.
E-mail address: sandra.oliveira1@campus.ul.pt (S. Oliveira).

https://doi.org/10.1016/j.mex.2023.102071
Received 15 September 2022; Accepted 7 February 2023
Available online 11 February 2023

Method details

Rainfall forecast is a critical component of a landslide or flash-flood early warning system since it allows to extend the lead time of reaction to a hazardous event. However, inaccurate rainfall forecasts may result in a reduction in the temporal and spatial accuracy of hazard predictions, compromising the early warning.

The present work was developed in the scope of a project that aims to create a landslide early warning soft technology prototype to improve community resilience and adaptation to environmental change in specific landslide prone areas of mainland Portugal. In one of the project tasks, we are receiving by FTP server and in text format, daily rainfall forecasts based on the regional Weather Research and Forecasting (WRF) model for 3 consecutive days. The Landslide Early Warning System (LEWS) will use these data to monitor different threshold levels for slope instability with 3 days of anticipation. Nevertheless, to ensure the accuracy of the forecasted rainfall, which will directly affect the results of the temporal and spatial prediction of landslides, it is necessary to estimate the difference between the forecasted and the real rainfall values. Therefore, in this work we developed a routine in R software that enables the validation of a 3-day rainfall forecast by comparison with the daily rainfall data recorded in 101 automatic meteorological stations available in mainland Portugal.

Base data

The base data comprises: (i) the rainfall forecast based on the weather regional forecasting model (WRF) for mainland Portugal, developed by Instituto Dom Luiz [1]; and (ii) daily rainfall recorded at each one of the 101 automatic meteorological stations (EMA) along Portugal's mainland, provided by the Portuguese Met-Office - Portuguese Institute for Sea and Atmosphere, I.P. (IPMA, IP).

The rainfall forecast based on the WRF model (WRF-R) is stored as 1 file per day, in *.txt format, with LAT/LON coordinates corresponding to the centroid of a grid of resolution of approximately 4 km (~0.03889º), in WGS84 projection system (EPSG: 4326). In the daily files, 4 values of forecasted rainfall are recorded (for the day of file name; and for the 3 consecutive days).

The daily rainfall registered at the 101 automatic meteorological stations (EMA, acronym in Portuguese) is provided in 1 file per year, in *.xlsx format. The 101 EMA stations are organized in a column and identified by the respective ID. Their location is given by a set of LAT/LON coordinates, in WGS84 projection system (EPSG: 4326). The daily rainfall measured at each EMA is disposed in rows (365 days/rows).

Pre-processing of base data

Concerning the WRF-R, we created a grid, in raster format, with the LAT/LON information available in the *.txt file as reference points (~centroids) of the cells. The location of the EMA stations, given as LAT/LON coordinates, were converted to a spatial points data frame and were superimposed in the raster files of WRF-R, to extract the forecast value (cell) corresponding to each EMA station. The creation of layers in raster format from the WRF-R files facilitates the subsequent intersection with EMAs stations, through a process of spatial analysis, for the automated retrieval of the cell values, from each WRF-R raster, that spatially overlap, i.e., which have the same location at the pixel level, with each EMA station. The processing was done with R software and tools packages [2], and the steps applied are detailed below

Import all files with forecast data, corresponding to the period under analysis

The procedure is done automatically, by defining the pattern (*txt), the path and folder where the WRF-R files are stored and importing them together to the R environment, within a list. All the files have the same 6 columns: LON, LAT, F1, F2, F3, F4. The names of the files were adjusted and simplified, and the column names of each file were changed to include the name of the corresponding file (FyearmonthdayFx). This step was necessary for further processing since a raster file will be created from each of these files and the corresponding name must be kept.

Convert each WRF-R table file to a spatial dataset

Based on a function created to apply the "coordinates" tool, the LON/LAT columns in each WRF-R file were used to create a spatial points data frame from each *.txt file. All the files have the same coordinates, and a common coordinate reference system (CRS) is defined (EPSG: 4326, WGS84), also using a user-defined function within a loop to apply the same parameters to all elements (spatial data frames) stored in the list. This intermediate step will facilitate the creation of the raster files afterwards.

Convert spatial points to raster file

A function was created to loop through all spatial points data frames and convert each of the columns (F1, F2, F3, F4) to a different raster file, keeping their names and storing them by day in a rasterstack. Simultaneously, the rasters are saved outside the R environment, in a folder specifically created for that purpose within the working folder, to allow for further visual inspection and mapping.

Import shapefile with coordinates of EMA stations and convert to spatial points

This shapefile includes the LON/LAT of the EMA stations, which will be used to extract the corresponding cells of the WRF-R set of rasters. The coordinate reference system of the shapefile is verified, and the unnecessary columns are removed.

Extract values of WRF-R rasters which intersect the EMA spatial points

From this step on, the only values of the WRF-R raster layers that are used are the 101 cell values that match the location of the EMAs. The ID of the cell extracted is initially retrieved, in case confirmation is needed, but it is later dropped to reduce the size and clean the resulting data frames that are used for further processing. The list of extracted raster values is converted to a set of data frames.

Combining data from WRF-R and EMA, based on sequence of days

Import excel file with EMA data, matching the period under analysis with WRF-R

The EMA files store the values of rainfall for each station in each day, separating the months in different sheets. The importation can be done by month, or for the whole file. Some EMA have negative values of rainfall, which correspond to missing data, which were replaced by NA (null values). The name of the first column of the EMA file was also changed. This column is not necessary for further processing, and it will be dropped since the order of the extracted points and the order of the EMA stations remains the same. The name of the other columns of the EMA file, which correspond to days, are changed to match the format of the forecast files (Dyearmonthday).

Create loop to pair columns of rainfall forecast (WRF-R) and EMA, in the corresponding sequence of days

The values of EMA and WRF-R were matched and joined (cbind function) by the column positions in each file, since all the files follow the same column and row order. This was done separately for each column of forecast (F1, F2, F3, F4), creating 4 different lists. For F1, the sequence starts in the first column of forecast and jumps every 4 columns in the next step (skipping all F2, F3 and F4), while retrieving all the successive columns of EMA, one by one, since the first column (first day). For F2, the sequence starts in the second column of forecast and jumps every 4 columns in the next step (skipping all F3, F4 and F1), while retrieving all the successive columns of EMA, one by one, since the second column (second day). A similar process was applied to F3 and F4.

Some adjustments were required to match the predicted (WRF-R) and observed (EMA) rainfall of the final days of the month. The last columns correspond to forecasted values for the first days of the following month. Therefore, the 3 first days of the successive month were previously added to the EMA files (e.g., February will have 31 days instead of 28 days). In the cbind sequences above, the number of columns iterated in the EMA files must be adjusted, adding 1 more for F2, 2 more for F3 and 3 more for F4. Another option would be to add 3 dummy columns in the EMAS values, with NA values, to allow the loop to run, but the matching values had to be disregarded. The loop of cbind was done for a month with 28 days (February 2015). If the size of monthly files is different, adjustments are required in the loop.

Create separate data frames by day

A loop was applied to retrieve from each list created in the previous step (F1, F2, F3, F4), the columns corresponding to each day and save them as a separate data frame. First, an empty list is created to aggregate all the data frames; then a loop is applied to iterate over all days of the forecast (February 2015 = 28) and extract from each list the column of the forecast of that day and the corresponding column of EMA day, sequentially. Each data frame will have the 4 columns of forecast and the corresponding matching days of the EMA files, intertwined. The data frames stored in the list do not have specific names, they can be called by number (=day in month).

Add columns with difference between EMA and WRF-R values in each data frame

Since all the data frames (1 per day) are stored in a list, different functions can be applied to all the data frames at the same time. In this case, a column with the original ID of EMA was added in the first position, to identify the stations when exporting the results afterwards. In addition, other columns were added with the difference between the values of rainfall measured at the EMA and the corresponding forecasted values, based on the position of the columns.

Export results as excel files

First, a new folder is created in the working folder, assigning a name related to the data analyzed. Then, each data frame stored in the list is exported individually to the new folder, into a separate excel sheet, but within a single excel file (e.g., ef201502.xlsx, 28 sheets).

Evaluation of WRF-R accuracy: forecast error metrics

Create new data frames where the data from each sheet (from excel file generated in 3.5.) corresponds to a column

The WRF-R and EMA values, as well as the difference of values between both variables for each one of the 4 days, are selected based on column names, using the grep function, or columns position in the data frame. The accuracy of the forecasted rainfall was estimated by computing the bias, the mean absolute error (MAE), the mean absolute percentage error (MAPE) and the root mean square error (RMSE), according to equations 1 to 4 [3]:

$$\text{Bias} = \frac{1}{N}\sum_{K=1}^{N}(P_K - O_K) \tag{1}$$

$$\text{MAE} = \frac{1}{N}\sum_{K=1}^{N}|P_K - O_K| \tag{2}$$

$$\text{MAPE} = \frac{MAE}{\frac{1}{N}\sum_{K=1}^{N} O_K} \tag{3}$$

$$\text{RMSE} = \sqrt{\frac{1}{N}\sum_{K=1}^{N}(P_K - O_K)^2} \tag{4}$$

Where O_K is the observed/real value P_K the predicted/forecasted value and N the number of pair values. The bias and MAE reflect the general deviation between predicted and observed values, allowing the identification of systematic errors. The MAPE calculates the relative measure of these errors regarding the mean observed values. The RMSE highlights the deviations between forecasted and real values due to the root mean [3]. To calculate the abovementioned metrics, the first column of each data frame is excluded since it corresponds to the EMA´s ID.

To automatize the procedure, the several data frames needed to compute each metric were aggregated in lists. The calculations were then carried out for each element within the lists based on looping functions, and the results were stored also as lists.

Moreover, if it is necessary to evaluate the forecast quality for each one of the 101 EMA stations, additional measurements can be performed, such as the contingency table [4], which is widely used to define if the forecasted values hit or miss the observed values.

Export forecast error metrics as excel file

The lists with the results obtained were exported to a single excel file where the error metrics are separated into sheets with the respective name. In each sheet, there are four columns, which correspond to the 4 forecasted days used in the calculations. An exception is made for the first sheet of the excel file, which records a single column with the EMA's ID (e.g., ef201502_Stats.xlsx).

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.

Data Availability

Data will be made available on request.

Acknowledgments

S.C. Oliveira was financed by the Portuguese Foundation for Science and Technology (FCT) I.P., under the framework of the project BeSafeSlide—Landslide EarlyWarning soft technology prototype to improve community resilience and adaptation to environmental change (PTDC/GES-AMB/30052/2017). SO was funded under the program of 'Stimulus of Scientific Employment— Individual Support', with the contract 2020.03873.CEECIND, and through national funds - UIDB/00295/2020- UIDP/00295/2020. AMR contribution was funded by the Helmholtz "Changing Earth" program. RT is funded by the Portuguese Foundation for Science and Technology (FCT) I.P./MCTES through national funds (PIDDAC) – UIDB/50019/2020.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.mex.2023.102071.

References

[1] Instituto Dom Luiz. (2007). WRF regional forecast for Portugal mainland. URL http://www.weather.ul.pt/.
[2] R Core TeamRA Language and Environment for Statistical Computing. R Foundation For Statistical Computing, Vienna, Austria, 2022 URL https://www.R-project.org/ .
[3] A.M. Ramos, R. Roca, P.M.M. Soares, A.M. Wilson, R.M. Trigo, F.M. Ralph, Uncertainty in different rainfall products in the case of two atmospheric river events, Environ. Res. Lett. 16 (4) (2021) 045012, doi:10.1088/1748-9326/abe25b.
[4] A. AghaKouchak, A. Mehran, Extended contingency table: performance metrics for satellite observations and climate model simulations, Water Resour. Res. 49 (2013) 7144–7149, doi:10.1002/wrcr.20498.