Accuracy in Precipitation Prediction

Robin Uhrich [*1] **Lilli Diederichs** [*2] **Mathias Neitzel** [*3] **Samuel Maier** [*4]

Abstract

In this report, we discuss the performance of weather prediction models provided by the German Weather Service (DWD). Over the past months, we've gathered weather forecasts from the DWD for Baden-Württemberg. As a reference, we use the measurements of the respective weather stations. We found that the mean error over all stations and forecasting horizons is diverting extremely, making general predictions for all stations difficult. However, the accuracy for predicting rain is higher in short-term predictions ($\approx 83\%$) than in those that predict further into the future (55% to 70%). We find that the location has a very small effect on the forecast in general but for local extreme rain events, we observed that the weather forecast drastically diverges from the reference data.

1. Introduction

The performance of a forecast is relevant in personal life and in agricultural planning. It has also become vital in measuring the risk of floods, which are becoming more and more prominent (Najibi & Devineni, 2018). Examples of regional significance are the Ahrtal Valley in 2021 (Pink S., 2023) and the recent floods in Germany around Christmas 2023 (DWD, 2023).

In this report, we discuss the performance of precipitation forecasting in Baden-Württemberg using the German Weather Service (DWD). We look at the performance of the 3-day and 10-day forecasts from 36 stations in terms of amount and timing. First, we calculate the mean and the mean absolute error. Second, we derive the accuracy to see how reliable the forecast is in timing. To further understand the predictions and the limitations of our data, we look at the impact of the location on the forecast. For that, we use a correlation matrix of the input features. Finally, we look at the station that was affected by the extreme precipitation over Christmas. (DWD, 2023).

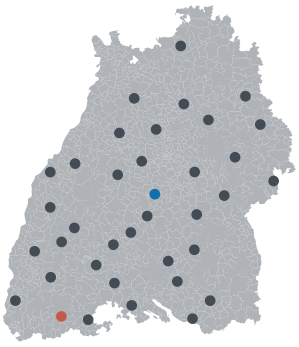

Figure 1. Map of Baden-Württemberg. Each selected station is represented by one dot on the map. The blue dot is Tübingen (which has no data), the red dot is Dachsberg-Wolpadingen

2. Data

Our analysis is based on data provided by the DWD. We collected forecasts on a daily basis for every weather station marked in Figure 1 from the Open-Data dataset. The dataset is divided into two categories:

- **Reference Precipitation**:
  Selected DWD weather stations measure precipitation with a rain gauge called rain[e]H3 (Lambrecht, 2023) of high accuracy ($\pm 0.001[mm/(mh)]$. These are accessible for recent and historical measurements through CDC dataset. We use these measurements as a reference to the forecast for each station.
- **Predictions**:
  The weather forecasts of all stations marked in Figure 1 are collected once a day at 00:10. The total precipitation is forecast every hour for the next 3 days, and every three hours for up to ten days.

We began our collection in 08-12-2023 and ended on 28-01-2024. Each 10-day forecast contains hourly predictions

[*]Equal contribution. University of Tübingen, Baden-Württemberg, Germany, MSc Machine Learning [1]matriculation number 6651884, robin.uhrich@student.uni-tuebingen.de [2]matriculation number 6638382, lilli.diederichs@gmail.com [3]matriculation number 4243096, mathias-neitzel@student.uni-tuebingen.de [4]matriculation number 4243096, sam.maier@student.uni-tuebingen.de.

The corresponding GitHub repository can be found here.

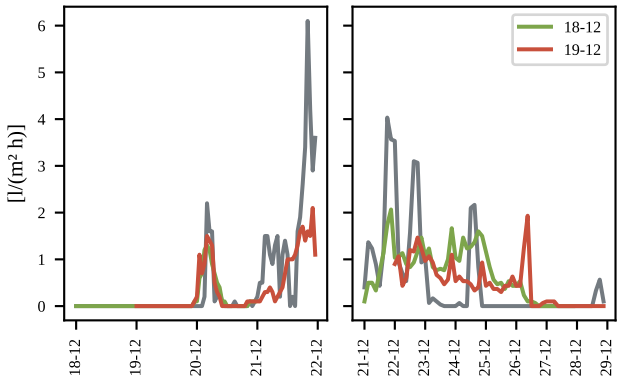

Figure 2. Raw Data for station in Dachsberg-Wolpadingen (red marked in Figure 1. Forecasts from 18-12-2023 and 19-12-2023 are in color. Reference data is in gray. The left plot is for the 3-day forecast, and the right plot is for 10-day forecast starting after day 3.

for the next 72 hours, and for the consecutive seven days, it contains predictions for every 3 hours, so 56 data points. Thus, we collected a total of $52 \cdot 36 \cdot (72 + 56) = 239.616$ samples. Because our reference data is limited to January 28, 2024, we have a total of 212258 samples. In Figure 2 you get an idea of what the data looks like for two queries.

3. Methods

As declared before, our data is split into two categories: forecasts $\hat{\mathbf{X}}$ and reference data $\mathbf{X}$. Where $\hat{\mathbf{X}}_{\Delta t,s,c,i}$ is the precipitation forecast at station s, queried at time c for time $c + \Delta t$, and $i \in \{1, 2\}$ differentiates between the 3-day and 10-day forecast, which describe one hour or three hours each. Respectively, $\mathbf{X}_{t,s}$ is the reference precipitation data at station s and at time t.

We denote N as the number of stations, T as the number of all possible timestamps with forecast and reference data, C is the number of queries executed, and Q is the number of time steps ahead of the forecast i is predicting.

3.1. Difference Measurement

To gain insight into the error of the forecast, we define two different metrics. The mean error in Equation 1 and the mean absolute error in Equation 2.

$$ME(\hat{\mathbf{X}}, \mathbf{X})_{\Delta t,i} = \frac{1}{NC} \sum_{s,c} \hat{\mathbf{X}}_{\Delta t,s,c,i} - \mathbf{X}_{c+\Delta t,s} \quad (1)$$

$$MAE(\hat{\mathbf{X}}, \mathbf{X})_{\Delta t,i} = \frac{1}{NC} \sum_{s,c} |\hat{\mathbf{X}}_{\Delta t,s,c,i} - \mathbf{X}_{c+\Delta t,s}| \quad (2)$$

For each day, we get multiple forecasts; in reverse, for each Δt we get 52 forecast differences, one for each collection day. By calculating the mean over Δt, we assume the forecast difference does not depend on the day. Secondly, we assume the ME and MAE to be independent of the location, such that we can find a mean deviation that applies to all stations over the collection time.

3.2. Accuracy

In order to estimate the performance of a given weather forecast, we use the accuracy $acc_{\Delta t,w}(\mathbf{X}, \hat{\mathbf{X}})$ as defined in Equation 3. This accuracy measure is based on the measure the DWD uses to track its quality (DWD, 2020), so that we can compare. We introduce a new variable w as the aggregation horizon, which defines how many samples into the future we take into account.

$$acc(\mathbf{X}, \hat{\mathbf{X}})_{\Delta t,w,i} = \frac{1}{NC} \sum_{s,c} B(\mathbf{X}_{[\Delta t,\Delta t+w),s}, \hat{\mathbf{X}}_{[\Delta t,\Delta t+w),s,c,i}) \quad (3)$$

$\mathbf{X}_{[\Delta t,\Delta t+w),s}$ is generated from the forecast data by accumulation. Specifically, we sum up every forecast, where the described duration of the forecast ($[\Delta t, \Delta t + w_{\text{data}})$) falls entirely into the duration of the accuracy $[\Delta t, \Delta t + w)$. For the 3-day forecast, we have w_{data} equal to one hour; for the 10-day forecast beyond it, we have w_{data} equal three hours. Our chosen w must be divisible by both; we consider multiple values for it.
With the binary mask B defined on two row vectors as in Equation 4, the first case is true negative, the second is true positive, and for false positive or false negative, we return 0.

$$B(\mathbf{x}, \hat{\mathbf{x}}) = \begin{cases} 1 & \forall x_i \in \mathbf{x},\ x_i = 0 \wedge \forall \hat{x}_i \in \hat{\mathbf{x}},\ \hat{x}_i = 0 \\ 1 & \exists x_i \in \mathbf{x},\ x_i > 0 \wedge \exists \hat{x}_i \in \hat{\mathbf{x}},\ \hat{x}_i > 0 \\ 0 & \end{cases} \quad (4)$$

The threshold to determine if we actually had rain or not is set to 0 as in (DWD, 2020)

4. Results

By observing the forecast and reference in Figure 2 we can clearly see that the forecast in orange and blue does not match the reference data in gray. We found a general absolute difference between forecast and reference data of 0.004 ± 0.392 for the 3-day forecast and 0.011 ± 0.353 for the 10-day forecast. Doing the analogous computation with MAE from Equation 2 we get 0.147 ± 0.363 for the 3-day forecast and 0.162 ± 0.314 for the 10-day forecast. The standard deviations in text and in Figure 2 are estimated with a bootstrap over 10^4 repetitions.

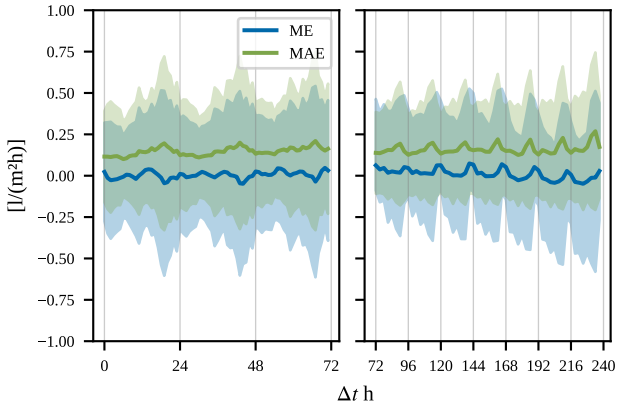

Figure 3. The mean difference between forecast precipitation and reference precipitation for depending on Δt in blue and the mean absolute difference in green. The left plot is for 3-day forecast and the right for 10-day forecast.

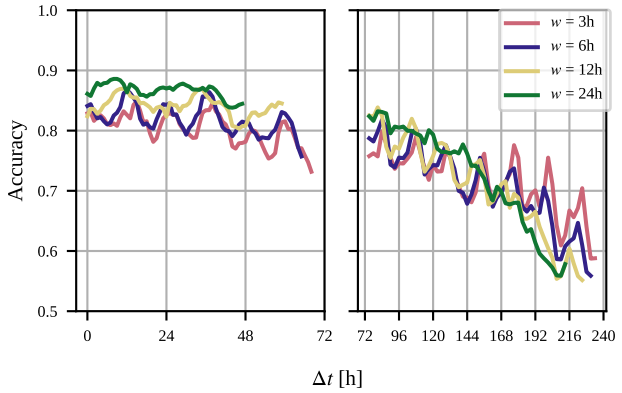

Figure 4. The accuracy for selected window sizes w and for forecasts Δt into the future. On the left hand side we display the 3-day forecast and on the right the 10-day forecast

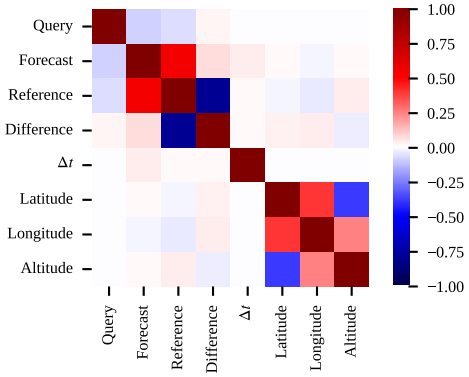

Figure 5. Correlation Matrix of input features of 3-day forecast

Because our estimated standard deviation is several magnitudes bigger than the mean value, the slight trend visible in Figure 2 for ME and MAE is definitely not significant. Nonetheless, the ME shows a daily pattern. Decomposing the oscillations with a Fourier transformation, we found the most dominant frequency of 13% amplitude at $3.0[1/day]$ and for the mean of the standard error, the most dominant frequency of 16% was at $1.0[1/day]$. The 10-forecast showed daily patterns as well, for the mean of 20% at $1.0[1/day]$ and the mean of the standard deviation of 16% $0.3[1/day]$.

Further, we are interested in the performance of predicting a rain event, no matter the amount. We use the accuracy proposed by (DWD, 2020) or (Frnda et al., 2022) and denoted for our use case in Equation 3. In Figure 4 we show the accuracy over Δt for different aggregation horizons. At this point, we were able to reproduce the accuracy over 12 hours for the 3-day forecast of $> 83\%$ released by the DWD in (DWD, 2020). Further, we can observe for the 3-day forecast that we have a positive correlation between the w and the accuracy. The 10-day forecast changes after $\Delta t \approx 5days$ from a positive to a negative correlation. To further investigate this change of slope, we looked at the error rates over Δt. We found a decreasing true positive rate while simultaneously increasing the false positive and false negative rates.

By averaging over all Δt, stations and queries as in Figure 3 we assume the difference between forecast and reference to be independently identically distributed with respect to the previously listed features. While this seems physically unrealistic, the assumption holds as we found no strong correlation with respect to the meta features: Δt location nor query date 5.

The calculated statistics for the mean difference or the mean accuracy have the strong disadvantage of under representing statistical outlayers which can be vital in certain situations. Especially in precipitation forecasts, it is crucial to predict floods. Therefore, we are interested in the quality of predicting such extreme events. As we can see in Figure 6 we have the accumulated sum of reference precipitation on the left and the average over queries from the 3-day forecast on the right.
Again, we can observe that the maximal prediction over all queries is on average larger than the reference. Even with the maximal forecast, it was not capable of predicting the 14-day heavy rain event in Dachsberg-Wolpadingen around Christmas (DWD, 2023). Resulting in a difference of $\approx 130[l/m^2]$ around the 29-12 for this specific station.

5. Discussion & Conclusion

Our first objective for this report was to determine the margin between forecast and reference. We found that the ME does oscillate around 0.004 for the 3-day forecast and around 0.011 for the 10-day forecast. The slight linear

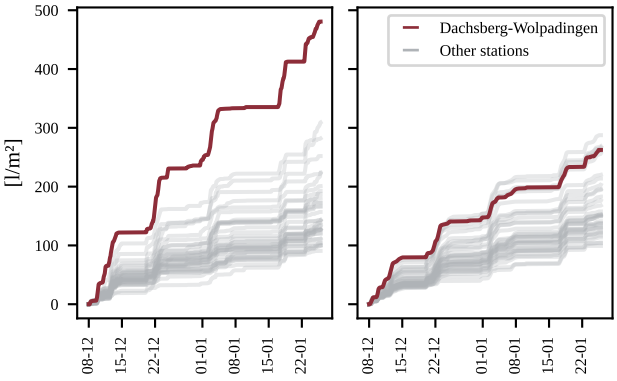

Figure 6. Cumulative sum of precipitation in December. Stations are in gray, except for Dachsberg-Wolpadingen marked red (For location, see figure 1). On the left we have the reference precipitation and on the right the accumulated mean forecast of the 3-day forecast

trend is negligible because the standard deviation per Δt is several magnitudes bigger than both trends. Also, the correlation matrix in 5 indicates that Δt does not have a strong correlation to the error. What we can account for is the present oscillation in the mean and standard deviation for the 3-day and 10-day forecasts for ME and MAE. As described in the results, we have two dominant frequencies in the Fourier transformation with a relative amplitude of at least 15%. Up to this point, we were not able to determine the source of this oscillation and therefore conclude that the source is probably not traceable with the features given in the dataset. We can only suspect that contributing factors are a strong correlation between MAE and the mean precipitation grouped over Δt or the scheduled daily query.

The second objective was to determine the accuracy of a weather forecast. Hereby, we concentrate on the accuracy grouped by the Δt. We are able to observe the same oscillations as in Figure 3. Discussing the accuracy for the 3-day forecast, we can state that with $w = 12$ we get a similar value to the released accuracy from (DWD, 2020) in 2018 of $\approx 83\%$. For a shorter window w the accuracy starts at lower values and decreases further over Δt. For the 10-day forecast on the other hand, the accuracy of the longer window w drops faster over Δt. As described in the results, this is due to a decrease in true positives and consequently an increase in false negatives. This is due to the binning. With increasing w the sensitivity for rain increases; therefore, the true positive rate is higher and the false negative rate is lower with $\Delta t = 0$. But for all observed true positive rates, we found convergence with maximum Δt and can conclude the faster dropping accuracy function depending of w and Δ. Therefore, we deduce that the 10-day forecast cannot be trusted even with a high binning horizon with respect to accuracy. Another factor could be presumed by the Figures 3 and 5, by which we presume an increasing sparsity of rain events but, on average, the same amount of discharge.

Regarding our final objective, we would like to know if the location of a station does have an impact on the quality. In pur previous analyses, we have seen almost no correlation between location features, like longitude, latitude, or height, to precipitation features. Also, the underlying data in Figure 3 does indicate a valid iid. assumption over the queried stations. But in December 2023, parts of Baden-Württemberg were affected by heavy rain events, like the station Dachsberg-Wolpadingen. As presented in Section 4 the analysis done in Figure 6 shows most stations that the reference distribution and the mean forecast distribution for the 3-day forecast do align. But for Dachsberg-Wolpadingen, according to our analysis, the highly accurate 3-day forecast was not able to predict the amount of rain at full expense. Therefore, the location could make a difference in certain situations and should not be handled with care.

To sum up our analysis, we came to the conclusion that weather is a highly complex topic, and quantifying the performance of a forecast has many ways to do it. Therefore, we cannot make definite claims within spatial and timely limitations of the data we collected, but we were able to show that our analysis does relate to analysis published by the DWD and goes beyond.

Acknowledgements

We would like to thank the DWD for providing access to the forecast and reference data. In addition, we would like to thank their Axel Kuschnerow for the quick answers to our questions. We also thank our tutor, Emilia Magnani, for her help and advice at any time.

Contribution Statement

Samuel Maier and Robin Uhrich jointly wrote the code to collect and preprocess the data. Together with Lilli Diederichs they performed the data analysis. In parallel, Mathias Neitzel helped formulate the analysis into formal statements. Lilli Diederichs was responsible for the visualizations. The text of the report was written jointly by all authors.

References

DWD. Qualität unserer Wettervorhersagen, 2020. URL https://www.dwd.de/DE/wetter/schon_gewusst/qualitaetvorhersage/qualitaetvorhersage_node.html#doc446682bodyText2. Accessed: 18-01-2024.

DWD, G. F. S. Deutschlandwetter im Dezember 2023, 2023. URL https://www.dwd.de/DE/presse/pressemitteilungen/DE/2023/20231229_deutschlandwetter_dezember2023_news.html?nn=16210. Accessed: 25-01-2024.

Frnda, J., Ďurica, M., Rozhon, J., Vojteková, M., Nedoma, J., and Martinek, R. Ecmwf short-term prediction accuracy improvement by deep learning. *Scientific Reports*, 12, 05 2022. doi: 10.1038/s41598-022-11936-9.

Lambrecht. rain[e] weighing precipitation sensor, 2023. URL https://www.lambrecht.net/en/products/precipitation/weighing-precipitation-sensor-rain-e. Accessed: 25-01-2024.

Najibi, N. and Devineni, N. Recent trends in the frequency and duration of global floods. *Earth System Dynamics*, 9(2):757–783, 2018. doi: 10.5194/esd-9-757-2018. URL https://esd.copernicus.org/articles/9/757/2018/.

Pink S., S. J. *Das Wetter ist politisch – Starkregen, Hochwasser und Flut vor der Bundestagswahl 2021*. Springer, 2023.