remote sensing

MDPI

Article

Satellite-Based Rainfall Retrieval: From Generalized Linear Models to Artificial Neural Networks

Lea Beusch *,†, Loris Foresti, Marco Gabella and Ulrich Hamann

MeteoSwiss, via ai Monti 146, 6605 Locarno-Monti, Switzerland; Loris.Foresti@meteoswiss.ch (L.F.); Marco.Gabella@meteoswiss.ch (M.G.); Ulrich.Hamann@meteoswiss.ch (U.H.)

* Correspondence: lea.beusch@env.ethz.ch; Tel.: +41-44-633-3624

† Current address: Institute for Atmospheric and Climate Science, ETHZ, 8092 Zürich, Switzerland.

Received: 16 May 2018; Accepted: 9 June 2018; Published: 13 June 2018

Abstract: In this study, we develop and compare satellite rainfall retrievals based on generalized linear models and artificial neural networks. Both approaches are used in classification mode in a first step to identify the precipitating areas (precipitation detection) and in regression mode in a second step to estimate the rainfall intensity at the ground (rain rate). The input predictors are geostationary satellite infrared (IR) brightness temperatures and Satellite Application Facility (SAF) nowcasting products which consist of cloud properties, such as cloud top height and cloud type. Additionally, a set of auxiliary location-describing input variables is employed. The output predictand is the ground-based instantaneous rain rate provided by the European-scale radar composite OPERA, that was additionally quality-controlled. We compare our results to a precipitation product which uses a single infrared (IR) channel for the rainfall retrieval. Specifically, we choose the operational PR-OBS-3 hydrology SAF product as a representative example for this type of approach. With generalized linear models, we show that we are able to substantially improve in terms of hits by considering more IR channels and cloud property predictors. Furthermore, we demonstrate the added value of using artificial neural networks to further improve prediction skill by additionally reducing false alarms. In the rain rate estimation, the indirect relationship between surface rain rates and the cloud properties measurable with geostationary satellites limit the skill of all models, which leads to smooth predictions close to the mean rainfall intensity. Probability matching is explored as a tool to recover higher order statistics to obtain a more realistic rain rate distribution.

Keywords: MSG SEVIRI; geostationary satellite; OPERA radar composite; rainfall; precipitation detection; rain rate estimation; generalized linear models; artificial neural networks

1. Introduction

Precipitation is a key component of the weather and climate system. Accurate rainfall measurements are essential for real-time nowcasting of severe weather and related hydrological applications, climatological precipitation studies, and to improve our understanding of the hydrological cycle, to name a few. Quantitative precipitation estimation (QPE) can be derived for a variety of temporal and spatial resolutions from rain gauges (e.g., [1,2]), ground-based weather radars (e.g., [3,4]), spaceborne weather radars (e.g., [5]), and satellite radiometers (e.g., [6–8]). Often, different data sources are combined to enhance the QPE accuracy, e.g., ground-based radar and gauges (e.g., [9,10]), ground-based radar and satellite radiometer (e.g., [11]), satellite radiometer and gauges (e.g. [7,12]), spaceborne radar and gauges (e.g., [13]), spaceborne radar and satellite radiometer (e.g., [14,15]), or spaceborne radar, satellite radiometer, and gauges (e.g., [16]). Each precipitation measurement instrument has its specific set of strengths and limitations.

Rain gauges directly measure surface precipitation, but, since they are point measurements, they lack representativity for regional-scale QPE [17,18]. This problem is further increased by the poor coverage of gauges over oceans and certain land regions [19]. Additional challenges include measurement errors, most importantly due to wind effects, but also caused by evaporation and wetting processes [20], with the largest uncertainty being observed for solid precipitation [21]. Weather radars are active remote sensing instruments that allow for monitoring the spatio-temporal evolution of precipitation systems over large areas. Radar-based QPE involves a complex data processing chain and is subject to several sources of uncertainty, among which there are the conversion of the measured reflectivity into rainfall rates, the attenuation of the signal, anomalous propagation, range degradation, spatial variability of the vertical profile of reflectivity, and residual clutter [22,23]. Satellite infrared (IR) and microwave (MW) radiometers, on the other hand, are passive remote sensing devices and QPE can be carried out based on the spectral information they provide. However, compared to active remote sensing instruments, the relationship between the spectral information and the rainfall is weaker. Nevertheless, satellite radiometer-based rainfall retrievals can deliver global-scale, spatially and temporally continuous, high-resolution precipitation products. Thus, they are most valuable in covering regions outside of radar domains or during radar-downtime [24]. They are especially helpful over oceans and in developing countries with no radar infrastructure, poor coverage, or limited resources to maintain the radar networks (e.g., [25,26]). Several comprehensive overviews treating different aspects of satellite-based QPE can be found in the literature [27–33].

In the following, "satellite-based precipitation" refers to rainfall that is derived solely from radiometers. Such precipitation products can be derived from low-level earth-orbiting satellites (LEO, typically at 350–850 km altitude; e.g., [7]), geostationary satellites (GEO, at about 36,000 km altitude; e.g., [34,35]), or a combination thereof (e.g., [8,36]). Since MW emission and scattering depend on the hydrometer size distribution, they are more directly related to rainfall intensity than cloud top temperature derived from IR measurements, making instantaneous rainfall retrievals with passive MW sensors on-board of LEOs attractive [37]. However, GEO satellites provide a much better temporal resolution rendering them more suitable for real-time precipitation monitoring [38]. Additionally, they give better estimations of daily and monthly precipitation accumulations. Adler et al. [39] were the first to combine the advantages of the two data sources by matching microwave and IR data to improve the algorithm of Arkin and Meisner [34].

The rainfall retrievals presented above are generally split into precipitation detection and rainfall rate estimation. Frequently, spectral information is taken as proxy for cloud top properties and conceptual understanding of rainfall processes is employed to determine the parametric relationships between satellite-derived information and precipitation patterns [40]. Precipitating areas are often distinguished from the non-precipitating ones using threshold tests for selected satellite channels and/or derived properties (e.g., [35,41–43]). Rainfall rates are then estimated by relating the satellite information to modelled or observed rain rates (e.g., [38,43,44]). If the goal is to obtain the best-possible performance in the rainfall retrieval rather than improving the conceptual understanding of the physical processes, machine learning approaches exploiting large numbers of input predictors can be helpful [40].

Also when using machine learning approaches, precipitation detection (e.g., [45,46]) and rain rate estimation (e.g., [47]) are often treated separately. Many studies can be found on such rainfall retrievals and we present a few in the following. The probably most widely known machine learning precipitation retrieval algorithm is PERSIANN [36]. PERSIANN, however, groups pixels according to cloud surface characteristics derived from GEO IR information with the use of an artificial neural network instead of explicitly identifying precipitating areas. For each group, a separate multivariate linear function mapping is employed to relate the input predictors to rain rates. LEO instantaneous rain rates are furthermore used to update the linear mapping network parameters. PERSIANN has been continuously developed e.g., by introducing a cloud classification system and thereby switching from pixel-by-pixel fitting of rain rates to cloud patches (PERSIANN-CCS, [48]). Recently, Tao et al. [49]

were able to significantly improve over PERSIANN-CCS in the delineation of precipitating areas by employing a deep neural network approach. Additional examples of neural networks being applied for satellite rainfall retrievals include neural network based fusions of LEO MW and GEO IR data [50] and precipitation detection with LEO MW observations [51]. Tapiador et al. [52] provide a general overview on satellite-based rainfall estimations by means of neural networks including a discussion about advantages as well as drawbacks of employing such a statistical-learning approach. In addition, the potential of other machine learning techniques such as random forests has been revealed for precipitation detection based on LEO MW [53] as well as for rainfall retrievals based on GEO IR and visible (VIS) channel input data [54]. Meyer et al. [40] compare several machine learning algorithms (random forests, artificial neural networks, and support vector machines) and show that the choice of machine learning algorithm only marginally affects the skill of the rainfall retrieval.

The main goal of this study is to explicitly investigate the added value of employing multivariate input predictors and machine learning methods for satellite-based rainfall retrievals. For this purpose, we develop and test linear and nonlinear statistical learning methods to estimate the occurrence and instantaneous rainfall intensity provided by the European Operational Program for Exchange of weather RAdar Information (OPERA) network. We use Meteosat Second Generation (MSG) GEO satellite brightness temperatures, cloud information provided by the Nowcasting Satellite Application Facility (NWC-SAF), and a set of auxiliary location-describing variables as predictors. The resulting satellite precipitation retrieval algorithms are compared to a single IR channel input predictor product, namely the operationally available instantaneous rain rate product of the Hydrology Satellite Application Facility (H-SAF). The additional contributions of this study reside in the combination of both brightness temperatures and NWC-SAF products for rainfall retrieval and the calibration of a GEO satellite precipitation retrieval using the European-scale OPERA radar composite.

The paper is organized as follows: Section 2 describes the used datasets and Section 3 the statistical learning methods. In Section 4, the results are presented by means of a representative verification as well as illustrative case studies. In Section 5, the results are discussed and set into a broader context. Finally, the conclusions are drawn in Section 6.

2. Data

The rainfall retrieval algorithm was derived and tested using satellite and radar data covering Europe during summer 2017 (mid-June to mid-August). In this section, we describe the employed satellite and radar datasets and the operational H-SAF product for instantaneous rainfall retrieval. Additionally, we discuss data quality aspects.

2.1. Satellite Data

The MSG Spinning Enhanced Visible and Infrared Imager (SEVIRI) instrument in full disk service performs a scan every 15 min with a spatial resolution of 3 km $\times$ 3 km at nadir for IR channels [55]. During the study period, this service was provided by Meteosat-10, which was located at 0 °E. We employed its Level 1b derived brightness temperatures and associated Level 2 NWC-SAF products [56,57] as satellite predictors for the rainfall retrieval. Specifically, we considered the following brightness temperature channels: WV 6.2 µm, WV 7.3 µm, IR 8.7 µm, IR 9.7 µm, IR 10.8 µm, IR 12.0 µm, IR 13.4 µm, as well as all the differences between them (corresponding to 21 additional variables), plus the following NWC-SAF products: cloud mask, cloud type, cloud top phase, cloud top temperature, cloud top pressure, and cloud top height. All of these predictors are available during both day- and nighttime and they were all parallax corrected based on the NWC-SAF cloud top height. The satellite data processing was carried out with the software package PyTroll [58].

In this study, rainfall predictions were made at the Europe-scale inside the NWC-SAF cloud mask. The map projection employed throughout the study shows the full prediction region as seen from the MSG SEVIRI in full disk service. It consists of 548 × 986 satellite pixels. It should be noted that the quality of the satellite products declines with increasing distance from the sub-satellite point due to lower spatial resolution and unfavorable viewing angles.

2.2. Auxiliary Variables

A range of auxiliary variables projected on the MSG full disk service grid were further included to characterize the spatial and temporal variability of the statistical relationship linking the satellite measurements to the ground-based radar observations. These additional variables are: the geographical coordinates (latitude/longitude), the local solar time, a land-water mask, and the altitude of topography.

2.3. Ground-Based Radar Reference

OPERA provides a European-scale radar composite, which is available every 15 min [4,59]. We employed OPERA surface rain rates re-projected on the MSG full disk service grid as a ground-based radar reference for the satellite rainfall retrieval. The OPERA composite offers a valuable continuous spatial and temporal coverage over land and it even partly extends over the ocean.

The quality of the OPERA radar composite increased substantially in 2017, e.g., it became considerably more successful in reproducing plausible precipitation frequencies across the Alps compared to summer 2016. In this study, precipitation frequency refers to the fraction of time slots during which the radar rain rates are $\geq$0.3 mm h^{-1} while the radar is in operation. The improved data quality in summer 2017 can partly be attributed to the inclusion of additional radars. The continuous improvement of the quality and coverage of this radar composite is very valuable but has the drawback of introducing temporal inhomogeneity in the data archive. To limit the effect of inhomogeneity while testing the potential of our statistical learning methods for satellite rainfall retrieval, we only used data from summer 2017.

Despite the improved data quality, some remaining issues needed to be addressed before the analyses. Radar-based QPE becomes more uncertain with increasing distance from the radar location due to beam broadening with distance and visibility effects, e.g., earth curvature and shielding by orographic features [60,61]. This often leads to an underestimation of precipitation frequency and intensity far from the radar. Additionally, residual ground and sea clutter, electromagnetic interferences, and wind farms may lead to falsely detected precipitation [62–64].

To eliminate uncertain radar measurements, we computed a map describing the frequency of precipitation $\geq$0.3 mm h^{-1} for summer 2017 (upper row in Figure 1). A radar composite mask was then derived by:

1. Removing all pixels that deviate too strongly from their neighborhood by:
    - Applying a 15 × 15 pixel running mean to the raw frequency map of Figure 1.
    - Subtracting this mean value from the actual frequency.
    - Subsequently removing all pixels with a deviation of $\geq$2.0% from the mean surrounding frequency, based on visual inspection. This step was effective in removing the locations affected by e.g., wind farms and radio-interferences.
2. Removing regions with implausibly low precipitation frequencies using a lower absolute frequency cut-off of 0.4%. Thereby, the locations were excluded where e.g., radar beams are blocked by topographic obstacles, the clutter rejection is too strong, or precipitation frequency is heavily underestimated at a large distance from the radar.
3. Eliminating the sea clutter affected region south of France that is connected to the Mistral winds.

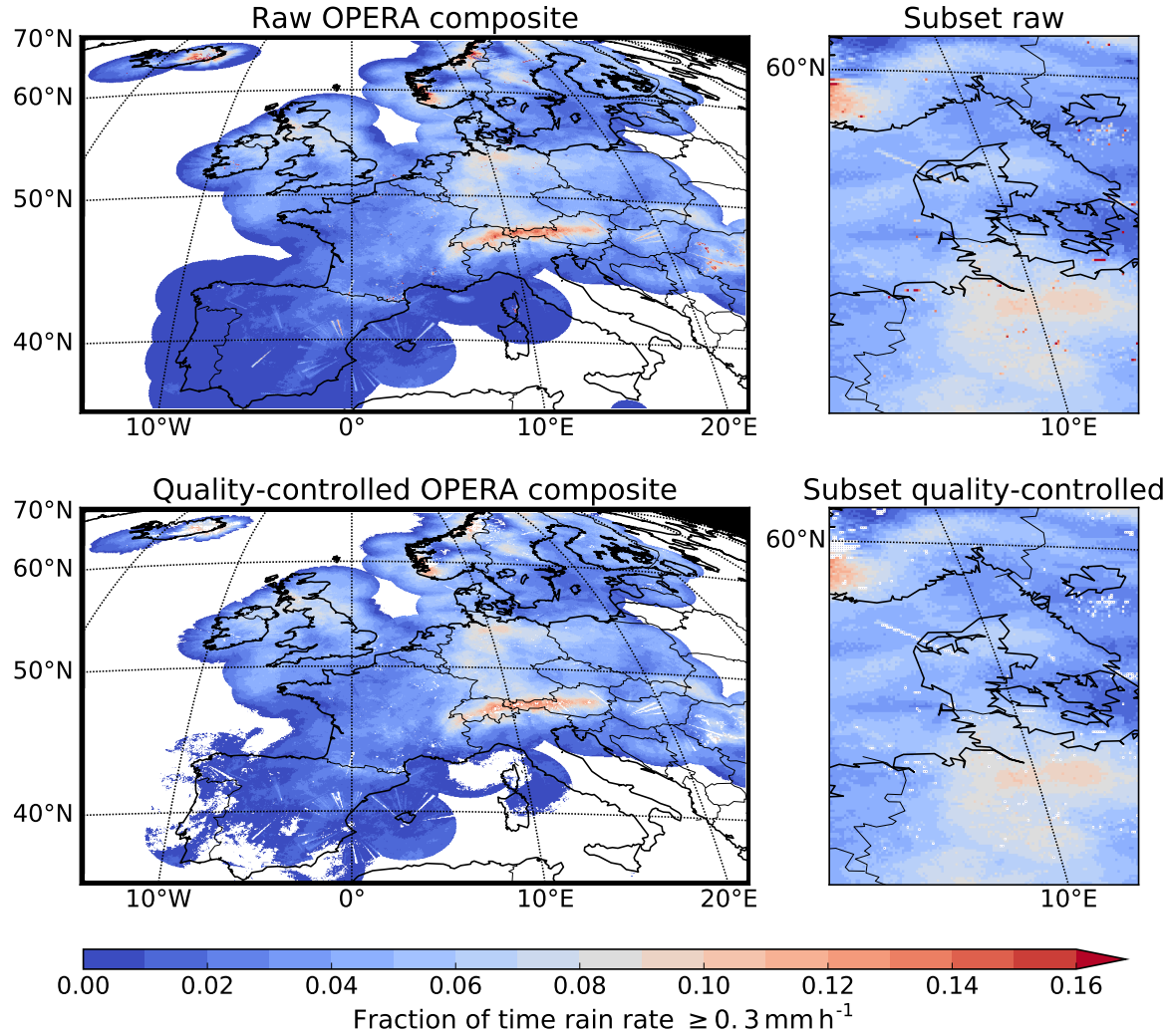

Figure 1. Maps depicting the fraction of time in summer 2017 during which the radar product is available and the rain rate is $\geq$0.3 mm h^{-1} for the raw OPERA composite (**top**) and the quality-controlled OPERA composite with unreliable regions masked out and thus shown in white (**bottom**). A zoomed-in version of the composite is provided on the right.

As shown, we applied a rigorous data quality control to avoid affecting the training of the statistical learning algorithms. As a consequence, we accepted also discarding a certain fraction of trustworthy precipitation measurements e.g., in dry regions like the Iberian Peninsula.

The lower row in Figure 1 shows the coverage of the OPERA composite after applying the radar mask. The majority of locations with low quality radar measurements were successfully removed. While the original OPERA composite covers 62.48% of the MSG pixels where predictions are carried out, the derived radar composite mask labels 54.95% of the MSG pixels of the prediction domain as reliable OPERA pixels.

An additional simple thresholding was applied on each individual radar image eliminating all rain rates $<$0.3 mm h^{-1} and $>$130 mm h^{-1}. It needs to be noted that, even after this procedure, individual radar images sometimes still contain residual clutter.

2.4. H-SAF Precipitation Product

The "EUMETSAT SAF on Support to Operational Hydrology and Water Management" or in short H-SAF provides satellite-based products for operational hydrological purposes [8,65]. In this study, we employed its PR-OBS-3 product, which is the instantaneous rain rate at the ground derived from MSG full disk service measurements "calibrated" with LEO MW measurements [8]. This product has the same spatial and temporal resolution as the MSG full disk service variables and we also parallax corrected it with PyTroll [58] based on the NWC-SAF cloud top height.

H-SAF derives the geolocated relationship between IR 10.8 µm brightness temperatures and MW rain rates from coincident IR and MW data at each grid box by a probabilistic histogram matching [8]. This relationship is updated whenever new data become available. Thus, a satisfactory correlation between the IR 10.8 µm channel and the rain rate at the ground is assumed. In the accompanying product user manual [66], it is stated that this is regarded acceptable for convective rain events, but less so for stratiform ones. Hence, the product shows a bias towards convective rain rates. Nevertheless, the authors of the product user manual claim that, at the moment, H-SAF is the only operationally available precipitation retrieval algorithm with a sufficient temporal resolution for nowcasting applications.

3. Methods

3.1. General Approach

The input predictors for the statistical learning algorithms for the rainfall retrieval consist of the SEVIRI IR channels, their differences, the NWC-SAF products, and the auxiliary variables described in the previous section, while the output predictand is the OPERA composite's instantaneous rain rate. In total, there are 51 predictors, which consist of 36 continuous and 15 dummy variables representing the categorical classes. Due to the collinearity of several input predictors, it was not possible to analyze their respective relevance [67]. Instead, we solely focused on the predictive power of the derived models, which should not be compromised unless the collinearity between the variable is subject to change.

The rainfall retrieval was performed in two steps:

1. Precipitation detection as a classification problem: All partly and fully cloudy pixels (according to the NWC-SAF cloud mask) were classified into either precipitating or non-precipitating pixels.
2. Rainfall intensity retrieval as a regression problem: The instantaneous surface rain rate was derived with a multivariate regression analysis on the previously detected precipitating pixels.

We investigated two different modelling setups. The first consisted of applying generalized linear models (GLM) with the classification being carried out by a logistic regression (LOGREG) and the regression by a linear regression (LINREG). The second considered artificial neural networks (ANN); specifically, the multilayer perceptron (MLP) was used both in classification and regression mode. The GLM and ANN models have been chosen because they are computationally fast, conceptually simple, and widely used in the literature. If there are substantial nonlinear data dependencies, the ANNs should be able to capture them and improve the prediction skill. If not, they are expected to provide similar skill as the GLM models. The precipitation detection and rain rate retrieval were performed using the same set of 51 predictors. The python scikit-learn package was employed for both the GLMs and the machine learning algorithms [68].

All models were compared to the operational instantaneous rain rate product of H-SAF to investigate the added value of considering additional satellite predictors by means of GLMs and using a nonlinear modeling approach with ANNs.

3.2. Data Splitting

Statistical learning requires splitting the dataset into a training, validation, and test set. The training set was used to derive the model parameters, such as the regression weights or the weights connecting the hidden neurons of the ANN. With the validation set, we selected the best hyper-parameters for the model, such as the number of hidden neurons and the regularization parameter. The test set was employed to estimate the generalization skill of the model on a new set of unseen data.

The time slots for training, validation, and testing were selected randomly with the additional constraints to contain a minimum of 4000 precipitating pixels inside the scene at hand and lying at least 45 min apart, i.e., the typical lifetime of a single cell thunderstorm. Hence, too highly correlated scenes were avoided. The training, validation, and test set consist of 800, 400 and 400 time slots during summer 2017, respectively.

To obtain balanced classes for the precipitation detection, we randomly selected 1000 precipitating and 1000 non-precipitating pixels inside the cloud mask for each time slot. For the rain rate retrieval, we substituted the non-precipitating pixels with additional precipitating ones, which resulted in 2000 randomly chosen precipitating pixels per time slot. For the validation and the test set, unbalanced datasets containing the actually observed distribution between precipitating and non-precipitating pixels were additionally generated for the precipitation detection using the same time slots. For this purpose, 10,000 samples were randomly drawn inside the cloud mask at each time slot. The resulting unbalanced validation and test set contain 7.4% and 7.6% precipitating pixels, respectively. In summary, the training, validation and test sets contain 1.6 mio, 800,000 and 800,000 samples. For the precipitation detection on unbalanced sets, the validation and test sets contain 4.0 mio samples each.

All categorical variables in the predictors described in Sections 2.1 and 2.2, such as e.g., cloud type, were converted into dummy variables, while the local solar time was expressed as a circular variable (sine and cosine) to eliminate the discontinuity at midnight. All 51 predictors were furthermore scaled to a mean of zero and a standard deviation of one based on the training set.

3.3. Verification Scores

For model validation and testing, a set of scores was considered. We refer to Appendix A for the meaning behind each score and for the respective equations.

The categorical scores for the classification problem were obtained by comparing the satellite-derived precipitating regions with the ground-based radar reference dataset. They all distinguish between events and non-events, which in our case refers to precipitating and non-precipitating pixels, respectively. We used Probability of Detection (POD), False Alarm Ratio (FAR), Probability of False Detection (POFD), Accuracy (ACC), and the Critical Success Index (CSI). Additionally, we calculated skill scores that measure the quality of a prediction system relative to a reference prediction, whereby a negative value indicates worse performance than the reference. In our case, the reference prediction is random chance that is represented by the "climatological" frequency of precipitation by pooling all the data samples in space and time. We computed the Gilbert Skill Score (GSS), the Heidke Skill Score (HSS), and the Hanssen-Kuipers discriminant (HK).

For the regression, the rain rate retrieved by satellite predictions was compared to the OPERA composite surface rain rate and the following continuous scores were considered: the Mean Error (ME), the Root Mean Squared Error (RMSE), the Pearson Correlation Coefficient (PCORR), the Spearman Rank Correlation (SCORR), and the Reduction of Variance skill score (RV). The RV skill score also measures skill compared to a "climatological" reference prediction, which is represented here by the observed sample variance.

3.4. Model Training and Hyper-Parameter Selection

In the GLM modelling chain, the default parameters of the scikit-learn package were taken for both the LOGREG and the LINREG with the exception of the solver used for the LOGREG, which was changed to the "Stochastic Average Gradient Descent" solver that is suitable for large datasets.

The ANN is a classical feedforward back-propagation MLP, which was trained using the gradient-based stochastic optimization algorithm "Adam" available in scikit-learn (see also [69]). The MLP was trained until convergence of the training set error. Overfitting and model complexity were controlled by optimizing the number of hidden layers, number of hidden neurons and the L_2 regularization parameter using the respective validation set. A logistic activation function was used to make a conceptual link between the linear logistic regression model and the nonlinear MLP model. Default values for the other MLP parameters were used, for instance a constant learning rate of 0.001 and a batch size of 200 points.

For the MLP classification, we employed the skill scores described in Appendix A.1: GSS, HSS, and HK. A grid search of parameter options was carried out to find the best combination of hyper-parameters based on the balanced validation set. Since this set contained 50% precipitating and 50% non-precipitating pixels, optimizing any one of these skill scores resulted in the same set of hyper-parameters. The best skill was achieved using an MLP with two hidden layers with 100 units each and a regularization parameter alpha equal to 10^{-7}.

In order to be able to carry out predictions for the actually observed unbalanced problem, the decision boundary, i.e., the probability threshold that is used to categorize model output into precipitating and non-precipitating pixels, needs to be adjusted with the help of the unbalanced validation set for the GLM and the ANN. We chose to optimize the GSS which automatically implies maximizing the HSS too (see Appendix A.1). This can be nicely observed in Figure 2. For the LOGREG model, the optimal threshold probability amounted to 0.835, and, for the MLP classifier, it was 0.870. It is furthermore evident that HK would be maximized at 0.500, since it considers both classes separately irrespective of their climatological frequency, which would lead to many false alarms in the unbalanced classification problem.

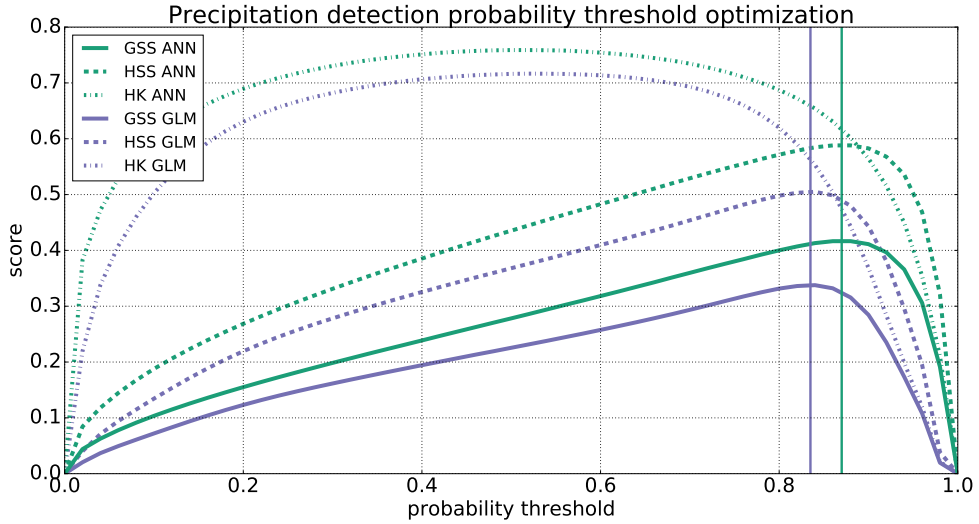

Figure 2. Gilbert Skill Score (GSS, solid line), Heidke Skill Score (HSS, dashed line), and Hanssen-Kuipers discriminant (HK, dash-dotted line) for varying probability thresholds in the precipitation detection in the unbalanced validation set. The artificial neural network (ANN) is given in green and the generalized linear model (GLM) in purple. The optimal thresholds according to the GSS are marked by vertical lines.

The hyper-parameters of the rainfall retrieval MLP regressor were optimized by maximizing the RV skill score described in Appendix A.2. However, the skill surface structure was not as clear in this case with several diverging parameter combinations resulting in similar skill for the validation set. We finally decided to use two hidden layers, 50 hidden units each, and a value of 10^{-2} for regularization. It is interesting to note that, while it was not possible to overfit on the classification problem, a certain magnitude of regularization was needed in the regression to avoid overfitting there.

3.5. Model Evaluation

The different methods were objectively compared by means of the representative test set. Additionally, single-scene case studies were carried out on time slots from the test set to analyze the realism of the predicted rainfall maps and the case-to-case variability of prediction skill. While the predicted rainfall is shown on the full map, the associated precipitation detection scores are computed—like in the representative test set—solely on trusted OPERA composite pixels (see Section 2.3) inside the NWC-SAF cloud mask (see Section 2.1). If each pixel inside the trusted OPERA composite were to be considered instead, the resulting single-scene skill would be higher since the pixels outside the cloud mask are by construction set to non-precipitating, which leads to many correct rejects.

To avoid punishing the regression algorithms for shortcomings in the precipitation detection, only pixels precipitating according to the quality-controlled OPERA composite were considered in the rain rate verification test set. On each one of these pixels, rain rates were predicted with the LINREG and the MLP regressor. For H-SAF, only a subgroup of the verification test set pixels was evaluated, namely the ones at which H-SAF detected precipitation, which amounts to 39.7% of the rain rate test set. In the case studies, rain rates were solely evaluated at the locations where both the trusted OPERA pixels and the classification algorithm detected rain. This implies that each algorithm was evaluated on a partly different set of pixels depending on the quality of the initial precipitation detection map.

3.6. Probability Matching

The LINREG minimizes the residual sum of squares between the OPERA measurements and the predicted rain rate. Similarly, the MLP regression minimizes the squared error loss function between them. Thus, strong deviations from the OPERA reference are heavily penalized and smooth predictions around the mean are favored. This leads to the best possible scores but not realistically looking precipitation fields. If visual consistency with radar fields and the ability to represent a more realistic rain rate distribution are valued higher than quantitative prediction skill, probability matching is a helpful tool. Probability matching carries out a nonlinear transformation to recover higher order statistics (such as variance and skewness) by means of empirical quantile matching of predicted rain rates with observed rain rates. In this study, the transformation from predicted rain rates into OPERA rain rates was derived on the validation set and then applied to the predictions resulting from the test set and the case studies.

3.7. Excluded Approaches

Over the course of the study, several additional approaches were tested but were deemed unsuitable for our purpose. In the following, we shortly mention the most noteworthy ones.

We investigated the potential of employing a range of textural features such as standard deviations computed within moving windows of 3×3, 15×15, and 20×20 pixels. The resulting negligible increase in skill did not justify the associated increase in computational cost. Thus, like Meyer et al. [70], we found no need to introduce structural parameters into our set of predictors.

The other excluded approach concerns the choice of the machine learning algorithm. As an alternative to the ANN, we additionally considered support vector machine (SVM) algorithms. However, they were found to be too computationally demanding in both training and more importantly also in prediction mode. The slow prediction speed originates from the need to compute the distances

between all support vectors and all prediction points. Thus, SVMs are not suitable to provide near-real-time satellite-derived precipitation maps over Europe. This matches with the findings of Meyer et al. [40], which point out the large computational costs associated with SVMs and slightly worse performance compared to ANNs.

4. Results

In this section, we analyze the results of the objective verification and support them with illustrative case studies in which we carry out European-scale predictions for a single time slot with each algorithm.

4.1. Objective Verification

The objective verification of the precipitation detection reveals that both GLM and ANN have higher skill than the H-SAF product when using the independent test set (Figure 3 circles). While the largest gain in performance is obtained by including the additional variables, applying machine learning further improves the predictions. It is interesting to note that, while the GLM already substantially increases the POD compared to H-SAF, the ANN succeeds much more in decreasing the FAR. Thus, adding input features mainly leads to the increase in hits while the machine learning is helpful in additionally reducing the number of false alarms. Only small improvements are observed in POFD and ACC since both of these scores depend on the number of correct rejects, which are large for each algorithm in this unbalanced classification problem. As expected, CSI and GSS show a similar improvement when increasing the number of input features and applying machine learning techniques, whereby the absolute magnitude of the GSS is smaller since it is additionally corrected for hits due to random chance. The HSS is higher than the GSS because the HSS measures accuracy adjusted for correct predictions obtained by random chance, i.e., it also takes correct rejects into account which are naturally large. To understand the behavior of the HK score, one needs to remember that it can be expressed as POD–POFD. Hence, the large improvement for the GLM compared to H-SAF is explained by its substantial increase in POD. The ANN's additional reduction of false alarms, however, does not strongly affect the HK score since POFD is dominated by the large number of correct rejects.

Instantaneous rain rate retrievals with GEO satellites are challenging and only result in little skill in any of the tested algorithms, which is due to the complex and indirect relationship between surface rain rates and satellite-observed variables (Table 1 test set). In terms of the RV skill score, the ANN performs best, closely followed by the GLM. As explained in Section 3.6, smooth predictions close to the mean observed rain rate (1.35 mm h^{-1}) are rewarded when deriving the GLM and the ANN. The resulting ANN manages to predict a maximum of 15.46 mm h^{-1}, while the GLM does not exceed 4.19 mm h^{-1}. Figure 4 shows the 2D histograms of observed against predicted rain rates, which further emphasizes the narrow prediction ranges of both models. Visually, the ANN seems to better capture the statistical relationship between the satellite input and the radar rain rates than the GLM. This is also confirmed by its slightly higher PCORR and SCORR (Table 1).

Probability matching recovers higher order statistics and thus covers the full range of observed values of 0.30–130.00 mm h^{-1} (Table 1). This procedure, however, leads to a decline in RV to below zero, indicating a worse performance than a "climatological" prediction represented by the observed sample variance would have. Additionally, the best linear fit between the observations and the predicted rain rates of the probability matched ANN deviates from the 1:1 line, i.e., a conditional bias is introduced (Figure 4). The linear relationship between the predicted rain rates and the observations, measured by PCORR, becomes less pronounced compared to the ANN and drops to the level of the GLM. The monotonic relationship, given by SCORR, in turn, remains constant since the ranks of the predictions do not change. In summary, despite the negative RV and the introduced conditional bias, there are several advantages to employing probability matched ANNs.

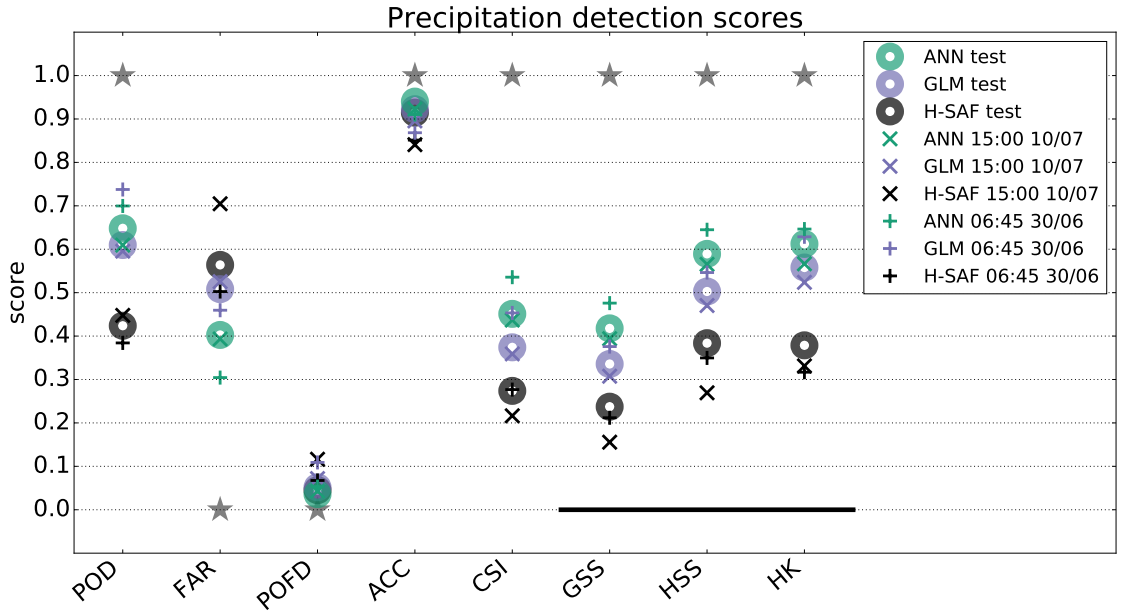

Figure 3. Precipitation detection verification scores for the test set (circles) and for the two case studies (15:00 UTC 10 July: convective, crosses; 06:45 UTC 30 June: stratiform, pluses) in summer 2017. The ANN is given in green, the GLM in purple, and H-SAF in black. The grey stars denote the perfect scores and the black horizontal line, the zero skill line.

Table 1. Rain rate estimation verification scores on the test set and the two case studies (15:00 UTC 10 July; 06:45 UTC 30 June) in summer 2017 for H-SAF, GLM, ANN, and the probability matched ANN (ANN PM). Min (max) pred refers to minimum (maximum) predicted rain rate and min (max) obs means minimum (maximum) rain rate observed by OPERA. # samples refers to the number of samples on which the predictions were evaluated, i.e., all the pixels in which both, the radar reference and the respective model detected rain. Mean Error (ME), Root Mean Squared Error (RMSE), min (max) pred, and min (max) obs are all given in mm h^{-1}, while Reduction of Variance skill score (RV), Pearson Correlation Coefficient (PCORR), Spearman Rank Correlation (SCORR), and # samples are unitless. Bold font is used to highlight the best performing algorithm for each score in each setup

	Test Set				**15:00 UTC 10 July—Convective**				**06:45 UTC 30 June—Stratiform**			
	H-SAF	**GLM**	**ANN**	**ANN PM**	**H-SAF**	**GLM**	**ANN**	**ANN PM**	**H-SAF**	**GLM**	**ANN**	**ANN PM**
ME	0.76	**−0.01**	−0.03	0.03	1.17	**−0.23**	−0.27	0.64	0.54	0.05	**0.04**	−0.32
RMSE	3.98	2.68	**2.64**	3.44	5.45	**4.61**	4.64	5.78	2.06	1.20	**1.13**	1.20
RV	−0.32	0.04	**0.07**	−0.58	−0.21	0.02	**0.05**	−0.47	−1.12	−0.01	**0.00**	−0.13
PCORR	0.18	0.21	**0.26**	0.21	0.16	0.17	0.24	**0.26**	**0.14**	0.08	0.09	0.08
SCORR	0.18	0.20	**0.27**	**0.27**	0.14	0.13	**0.21**	**0.21**	**0.23**	0.09	0.15	0.15
Min pred	**0.30**	**0.30**	**0.30**	**0.30**	**0.30**	0.59	0.66	**0.30**	**0.30**	0.61	0.65	**0.30**
Min obs	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30	0.30
Max pred	29.45	4.19	15.46	**129.68**	14.09	3.92	12.44	**114.62**	**7.18**	2.81	2.57	4.13
Max obs	129.70	129.70	129.70	129.70	110.52	110.52	110.52	110.52	52.57	52.57	52.57	52.57
# samples (×10,000)	31.76	80.00	80.00	80.00	0.65	0.86	0.88	0.88	0.78	1.50	1.43	1.43

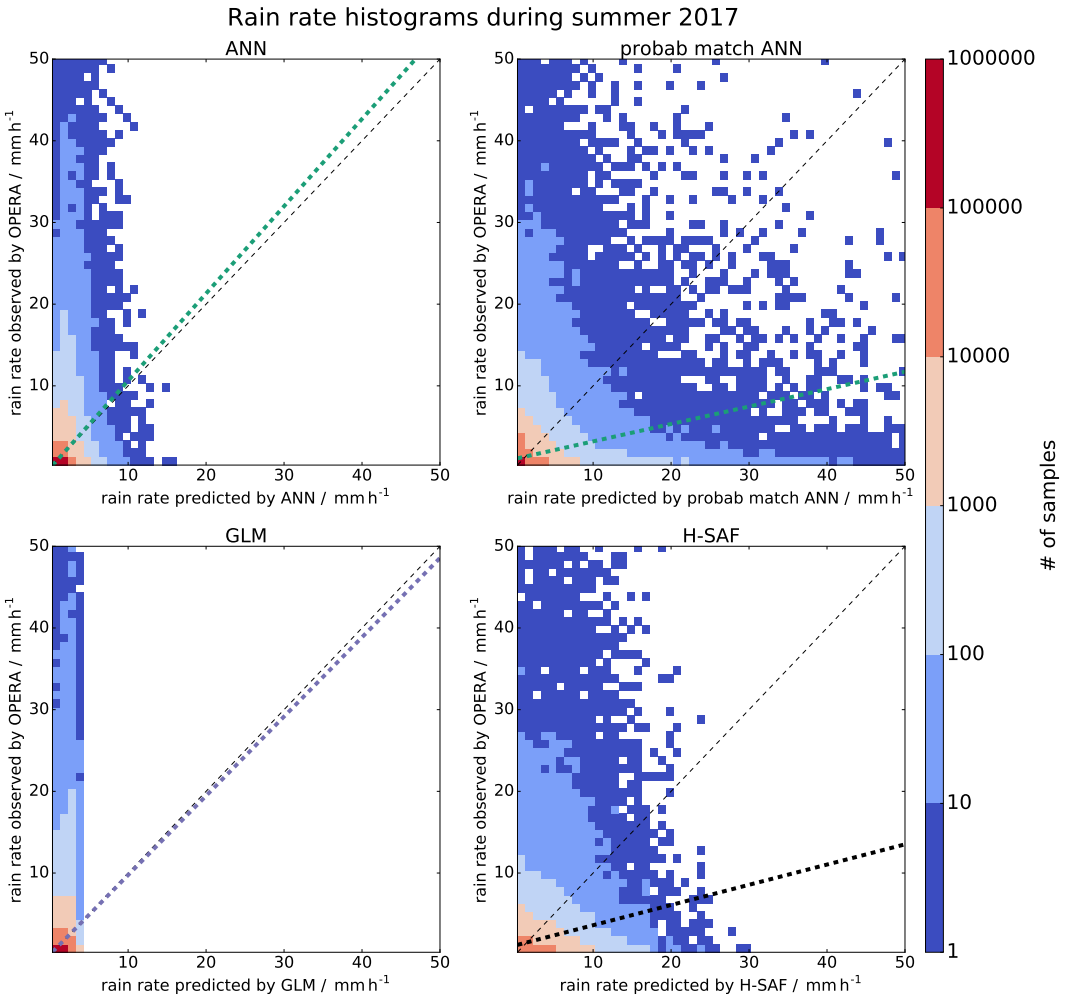

Figure 4. 2D histograms showing the observed rain rates by OPERA in mm h^{-1} on the y-axis and the predicted rain rates by each model in mm h^{-1} on the x-axis (ANN, probability matched ANN, GLM, and H-SAF). The number of samples in each 1 mm h^{-1} $\times$1 mm h^{-1} bin starting at 0.3 mm h^{-1} is indicated by a logarithmic color scale. The diagonal thin dashed black lines depict the perfect 1:1 linear fit. The thick dashed lines mark the best linear fit for each model, with the ANN given in green, the GLM in purple, and H-SAF in black. More information on each model's performance can be found in the test set section of Table 1.

H-SAF shows the lowest correlations of any of the tested approaches and is also faced with a RV skill score below zero (Table 1). It predicts rain rates up to 29.45 mm h^{-1}. Hence, it is outperformed by each one of our algorithms in both precipitation detection and rain rate estimation. While a clear distinction in model performances is seen in the precipitation detection, the rain rate estimation performance varies less between the models. It should be pointed out that, while the GLM and the ANN were trained on OPERA data, H-SAF was calibrated on MW measurements. To quantify the importance of this effect, we trained an additional GLM with the same input variable as H-SAF, namely the IR 10.8 µm channel. The model we obtained performs very similarly to H-SAF in terms of skill (not shown). Thus, we conclude that the improvement of our multivariate GLM compared to H-SAF can be predominantly attributed to the additional variables, which are considered without needing to take the different "calibration" datasets into account.

4.2. Illustrative Case Studies

Case studies allow us to put the conclusions drawn from the objective verification in the context of the actually desired single-scene rainfall prediction maps. While our first case study depicts widespread convection over the European continent, the second one contains large, mostly stratiform, rainbands, thus making it possible to compare the algorithms' performances qualitatively in different weather regimes.

4.2.1. Convective Case—15:00 UTC 10 July 2017

In terms of precipitation detection, the skill in the convective case is slightly below average for the GLM and the ANN while it is very low for H-SAF (Figure 3 crosses). This is almost exclusively attributable to the H-SAF's large amount of false alarms, which is reflected in its large FAR. This strong H-SAF outlier in terms of false alarm performance highlights the big case-to-case variability of the scores and underlines that the average performance derived from the test set is not representative for every single scene. However, it should also be noted that we specifically chose to show an extreme case of a false alarm outlier performance here.

When visually comparing H-SAF to the GLM, a strong decrease in false alarms and increase in hits is observed in the GLM, with the improvements largely limited to continental Europe and most progress being made in northeastern Europe (Figure 5). The high opaque clouds in northeastern Europe (not shown) were mistaken as precipitating by H-SAF, which is a natural consequence of their cold IR 10.8 µm brightness temperatures. In this case study, the ANN manages to further reduce the false alarms, while the amount of hits remains similar to the GLM. Nevertheless, it is arguable that, visually, the ANN also outperforms the GLM in terms of hits since it is the only algorithm able to detect precipitation in the northern parts of Great Britain. While the ANN does not always identify the full extent of the precipitating areas, it is most successful at identifying regions faced with rainfall in this case study.

The rain rate estimation performance of each product is qualitatively similar to the objective verification with generally smaller correlations and larger errors (Table 1 convective). The map shown in Figure 5 highlights the main issues of the rain rate retrieval. None of the satellite-based methods are able to capture the fine-scale precipitation structures displayed by the ground-based radar network. The GLM's smooth predictions around the mean make it visually most different from the radar composite. The ANN, on the other hand, is able to reproduce finer spatial structures deviating more from the mean while still outperforming the GLM in terms of scores. Only probability matching allows for an overall similar rain rate distribution compared to the radar composite (Figure 6). However, the structures remain less fine-scale than in the radar composite and, due to the double penalty effect, ME and RMSE increase strongly while RV drops below zero (Table 1).

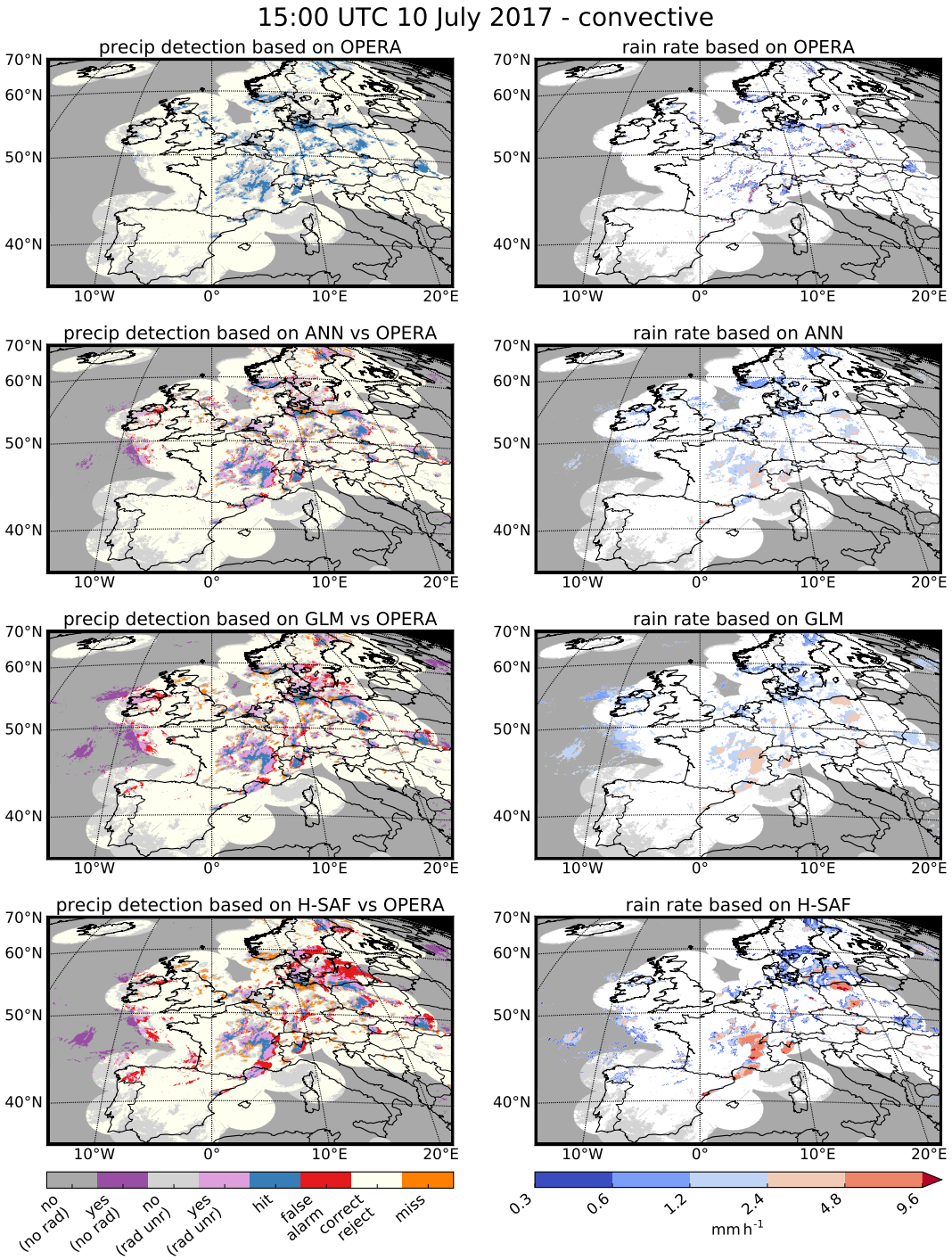

Figure 5. Case study at 15:00 UTC 10 July 2017 of precipitation detection (**left**) and rain rate estimation (**right**). The first row consists of the OPERA observations. The subsequent rows show the ANN, GLM, and H-SAF predictions. The precipitation detection is evaluated against the OPERA radar (rad) reference. "Yes" means precipitation, "no" means no precipitation according to the system at hand. On reliable OPERA pixels, the predictions are referred to by their contingency table names: hit, false alarm, correct reject, and miss. Predictions on unreliable radar pixels (rad unr, i.e., on the permanent radar mask derived in Section 2.3 and where the radar rain rate is the range of 0.1–0.3 mm hh^{-1} or >130.0 mm hh^{-1} in this scene) are distinguished from pixels where no radar (no rad) is available. In the rain rate plots, the regions where no radar is available and the permanent radar mask are also marked.

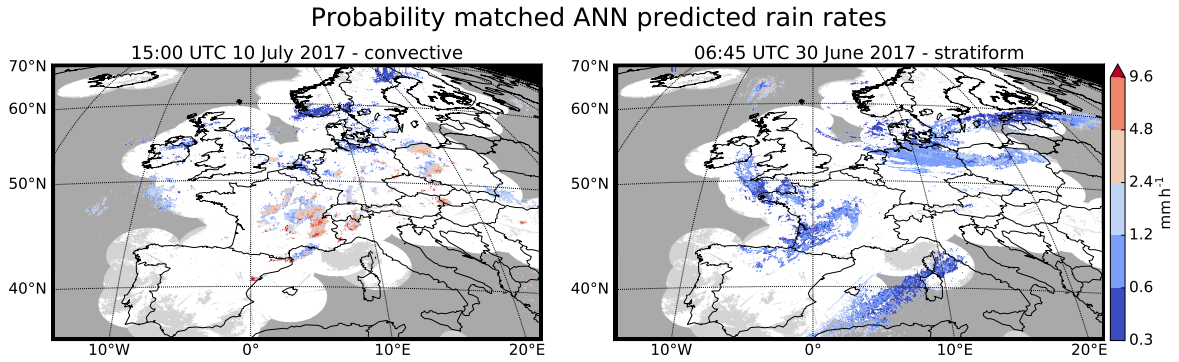

Figure 6. ANN rain rate retrieval product after probability matching of the predicted rain rates. On the left for 15:00 UTC 10 July 2017 (see also Figure 5) and on the right for 06:45 UTC 30 June 2017 (see also Figure 7).

4.2.2. Stratiform Case—6:45 UTC 30 June 2017

In the precipitation detection of the stratiform rainband case, an unusually high skill is observed for both the GLM and the ANN (Figure 3 pluses). On the one hand, this reveals that our algorithms, similar to H-SAF, are faced with large case-to-case variability in performance. It thus emphasizes the importance of using the test set containing 400 time slots to estimate the average generalization skill of our algorithms instead of solely relying on single case studies. On the other hand, Figure 7 shows that, in this specific case, the verification scores profit from the presence of large regions in which no scores are computed because the radar composite depicts rain rates between 0.1–0.3 mm h^{-1} (which we do not consider due to reliability issues). There is an uncommonly large buffer separating the precipitating areas from the non-precipitating ones resulting in lower FAR than on average for each one of the algorithms (Figure 3). Similar to the objective verification, the GLM mostly increases the number of hits while the ANN adds additional value by further decreasing the number of false alarms. A logical consequence of H-SAF only considering the IR 10.8 µm channel for the precipitation retrieval is its inferior capability to detect stratiform precipitation originating from low- and mid-level clouds. This is especially visible in the large rainband structure extending from Ireland down into France and northern Spain where the cloud top height is in the 6–8 km range (not shown) and which is almost entirely missed by H-SAF (Figure 7).

In this stratiform case (Table 1 stratiform), the performance of the rain rate retrieval in terms of RV skill is below average for every algorithm, except for the probability matched ANN. The same conclusion is reached in terms of correlations apart from the H-SAF SCORR, which is unusually high. All algorithms predict lower rain rates and thus a smaller rain rate range than they did in the convective case and in the test set, which explains the smaller RMSE. Even probability matching only results in a maximum rain rate of 4.16 mm h^{-1} inside the quality-controlled radar mask, which is far below the maximum radar observed 52.6 mm h^{-1}. The resulting transformed predictions are shown in Figure 6.

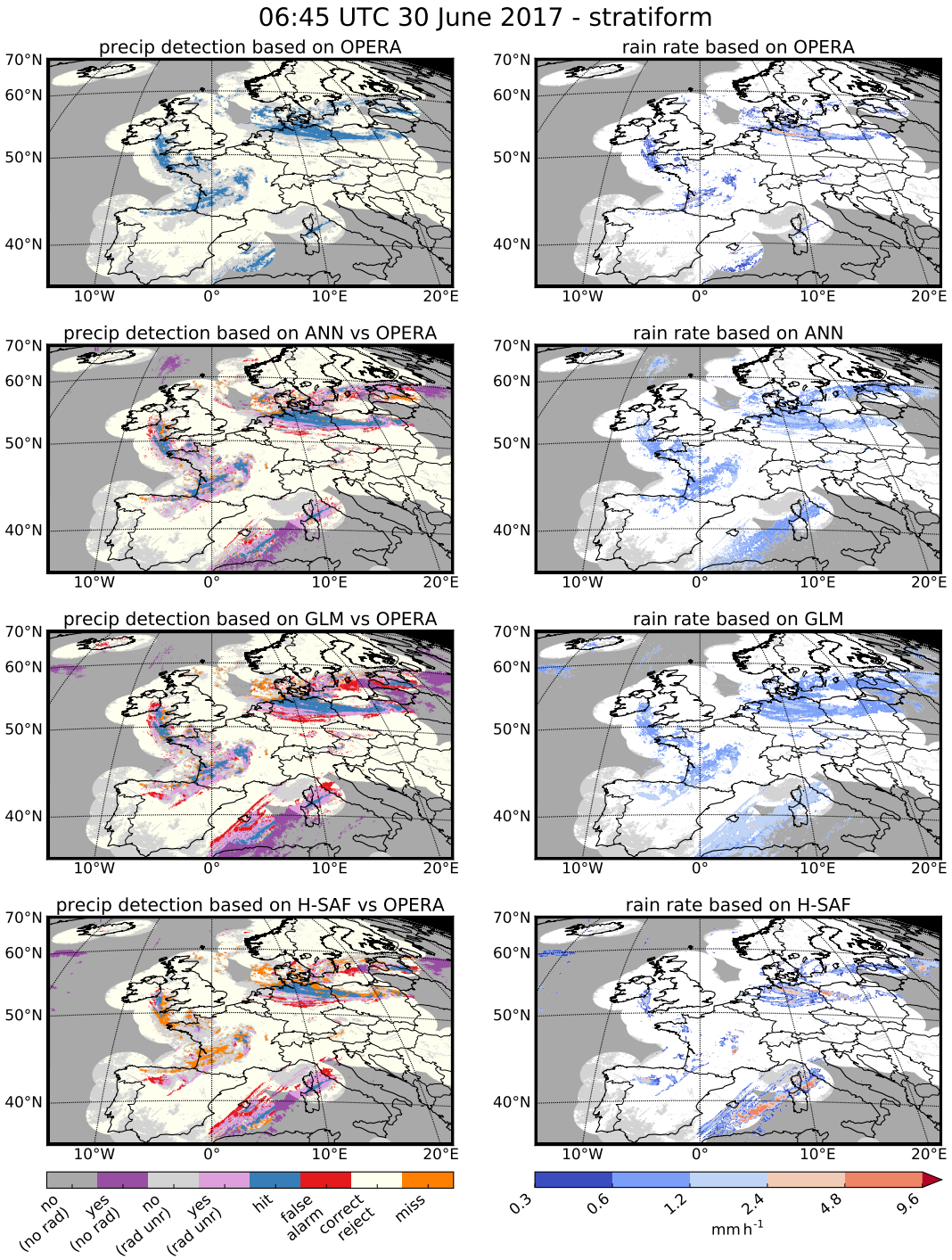

Figure 7. Same as Figure 5 but for 06:45 UTC 30 June 2017.

5. Discussion

In this section, we discuss general features uncovered with case study maps, strengths and limitations of our rainfall retrievals, and possible ways forward together with the associated challenges.

Visual inspection of many precipitation detection maps including Figures 5 and 7 revealed that the GLM and the ANN do not always perform best in the same regions. For example, some precipitating areas that are captured by the GLM, such as the rain over Latvia in Figure 7, fail to be predicted by the ANN and vice versa e.g., for the scattered precipitation over Great Britain in Figure 5.

Furthermore, the maps indicate a decrease in performance of the algorithms at high latitudes, especially to the northeast and over Iceland (not shown). Computing the test set scores for only the pixels originating from these regions confirms our visual impression. This is probably caused by the unfavorable viewing geometry of the satellite at these locations and the stronger deformation of the OPERA radar data when projected onto the satellite grid at high latitudes. Potentially, this issue could be somewhat diminished by replacing the predictors from the MSG full disk service with MSG rapid scan ones, since the rapid scan satellite has a more favorable viewing geometry due to its location at 9.5°E, which is more centered above the OPERA domain. Additionally, the mountainous regions in Iceland imply difficult conditions for the radar-based QPE, which is demonstrated by the large number of discarded clutter pixels (see radar mask in Figure 1). Nevertheless, some residual clutter remains in the dataset, which negatively affects the ANN training and predictions over Iceland.

Another issue is the tendency of our rainfall retrievals to also predict high rain rates at the edge of precipitation systems (see, e.g., Figures 5 and 7). H-SAF, on the other hand, seems to struggle less with this. One possible solution could be to integrate as additional predictor a variable characterizing the "dry drift", i.e., the distance from the edge of the closest surrounding dry area (e.g., [71]).

One of the biggest challenges of the current study is the training of the rainfall retrieval algorithms using pixel-based instantaneous rain rates. In fact, it is widely known that correlations between predictions and observations tend to improve with increasing spatial (e.g., [38]) and temporal aggregation (e.g., [36,40,47,54]). We nevertheless decided to focus solely on pixel-based instantaneous rain rates here because they are especially valuable for real-time nowcasting applications, as they can e.g., be used to fill in gaps during radar downtime and cover regions with poor or no radar coverage. Nonetheless, one should keep in mind that our product is clearly better suited for identifying precipitating regions than for instantaneous QPE.

Several other studies (e.g., [13,45,46]) showed that, during the daytime, satellite-based rainfall retrievals benefit from additionally considering VIS channels. However, when taking VIS channels into account, two separate algorithms need to be used for day- and nighttime with the nighttime algorithm exhibiting lower skill. An IR-only approach, as employed in this study, on the other hand, results in a single algorithm whose skill does not depend on whether it is day or night, which can be argued to be advantageous.

It would furthermore be valuable to extend the training set to represent a longer time period that also includes winter months, and thus learn seasonal dependencies. Currently, this is complicated by the ongoing efforts to improve the OPERA composite, which creates temporal data inhomogeneity. Another challenge is given by the seasonality of the quality of both radar and satellite products since each of them faces more uncertainties during the cold season. Nonetheless, the work currently invested by EUMETSAT and OPERA results in datasets with increasing quality to train our algorithms, which will in turn likely be reflected in an increase in prediction skill.

6. Conclusions

In this study, we demonstrated, visually as well as quantitatively, the added value of multivariate input predictors and machine learning methods for geostationary satellite-based instantaneous rainfall retrievals. We trained generalized linear models (GLM) and artificial neural networks (ANN) using infrared brightness temperature information, NWC-SAF cloud property products, and auxiliary variables as input predictors. The target variable for the statistical learning models is represented by rain rates provided by the quality-controlled ground-based European radar composite OPERA, which was also used as the reference for verification. The resulting predictions were compared to an operational single-input-predictor-based instantaneous precipitation product, namely PR-OBS-3 of H-SAF.

The rainfall retrieval process is divided into two steps for which individual models were trained: precipitation detection (a classification problem) and rain rate estimation (a regression problem). The additional input predictors lead to a strong increase in Probability of Detection for the GLM

compared to H-SAF. Taking nonlinearities into account with the ANN mostly adds value by further reducing the false alarms compared to the GLM. Despite large case-to-case variability in skill, both the GLM and the ANN outperform H-SAF in all the tested weather situations.

The indirect relationship between geostationary satellite information and surface precipitation limits the skill of the rain rate estimation. While we achieve superior scores compared to H-SAF, the H-SAF product is better at creating plausible looking precipitation fields. In fact, the training procedure of our algorithms penalizes large deviations from observations heavily, which leads to smooth prediction around the mean rain rate. The ANN is only slightly superior to the GLM in terms of scores but is less smooth and creates more fine-scale structures inside the precipitating areas. To improve the visual realism and the representation of higher order statistics (such as variance and skewness) of the predictions, we transformed the predicted rain rates based on probability matching. This, however, results in inferior verification scores.

The performance of the ANN is promising and motivates further research to improve its rainfall product. This could be achieved by using rapid scan MSG data to have a more favorable viewing geometry and extending the training and testing sets to include longer time periods and also the cold season.

Author Contributions: U.H. had the initial idea for the study and provided the data. L.B. and L.F. set up the learning problem and designed the experiments. L.B. carried out the analysis with machine learning support from L.F., radar support from M.G., and satellite support from U.H. L.B. wrote the article. All authors commented on the manuscript and analyses.

Funding: No external funding was received by L.B., M.G. and U.H. L.F. was supported by the Swiss National Science Foundation Ambizione project "Precipitation attractor from radar and satellite data archives and implications for seamless very short-term forecasting" (PZ00P2_161316).

Acknowledgments: We would like to thank MeteoSwiss for giving us the opportunity to carry out this study. Additionally, we are thankful to EUMETSAT, OPERA, the NWC-SAF and H-SAF for providing us with their data.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

Institution related abbreviations:

H-SAF	EUMETSAT Satellite Application Facility on Support to Operational Hydrology and Water Management; or in short: Hydrology Satellite Application Facility
OPERA	Operational Program for Exchange of weather RAdar Information
NWC-SAF	Nowcasting Satellite Application Facility

Measurement related abbreviations:

GEO	Geostationary satellite
IR	Infrared
LEO	Low-level earth orbiting satellite
MSG	Meteosat second generation
SEVIRI	Spinning enhanced visible and infrared imager
VIS	Visible

Model related abbreviations:

ANN	Artificial neural network
GLM	Generalized linear model
LINREG	Linear regression
LOGREG	Logistic regression
MLP	Multilayer perceptron
SVM	Support vector machine

Scores related abbreviations:

ACC	Accuracy
CSI	Critical Success Index
FAR	False Alarm Ratio
GSS	Gilbert Skill Score
HK	Hanssen–Kuipers discriminant
HSS	Heidke Skill Score
ME	Mean Error
PCORR	Pearson Correlation Coefficient
POD	Probability of Detection
POFD	Probability of False Detection
RMSE	Root Mean Squared Error
RV	Reduction of Variance skill score
SCORR	Spearman Rank Correlation

Appendix A

Appendix A.1. Categorical Scores for the Precipitation Detection

The categorical scores explained in the following are all based on the contingency table shown in Table A1. More thorough descriptions can be found in the book of Hogan and Mason [72].

Table A1. Contingency table introducing the abbreviations used to compute the scores.

		Observations	
		Yes	**No**
Predictions	**Yes**	hits (H)	fase alarms (F)
	No	misses (M)	correct rejects (R)

The Probability of Detection (POD, also called hit rate) measures the proportion of the correctly predicted events among the total number of observed events:

$$POD = \frac{H}{H+M}.$$

The False Alarm Ratio (FAR) depicts the proportion of false alarms in all predictions, i.e., it is conditioned on the predictions:

$$FAR = \frac{F}{F+H}.$$

The Probability of False Detection (POFD, also called false alarm rate) shows the proportion of non-events predicted as events, i.e., it is conditioned on the observations:

$$POFD = \frac{F}{F+R}.$$

The Accuracy (ACC, also called fraction correct) is the proportion of correct predictions in all predictions:

$$ACC = \frac{H+R}{H+F+M+R}.$$

The Critical Success Index (CSI, also called Threat Score) indicates the proportion of hits in all predicted (i.e., hits and false alarms) and missed events:

$$CSI = \frac{H}{H+F+M}.$$

The Gilbert Skill Score (GSS, also called Equitable Threat Score) is equal to the CSI adjusted by hits due to random chance. It is sensitive to hits and the error source cannot be determined because misses and false alarms are penalized in the same way:

$$GSS = \frac{H - H_r}{H + F + M - H_r} \quad \text{with the hits expected by random chance: } H_r = \frac{(H + F) \cdot (H + M)}{H + F + M + R}.$$

The Heidke Skill Score (HSS, also called Cohen's k) measures the proportion of correct predictions, i.e., the accuracy, after removing the predictions that are correct due to random chance. It can be regarded as a rescaled version of the GSS to obtain a linear score and can be expressed as: $HSS = \frac{2 \cdot GSS}{1 + GSS}$:

$$HSS = \frac{(H + R) - (H_r + R_r)}{(H + F + M + R) - (H_r + R_r)} \quad \text{with } H_r \text{ like in the GSS and } R_r = \frac{(R + F) \cdot (R + M)}{H + F + M + R}.$$

The Hanssen–Kuipers discriminant (HK, also called Peirce Skill Score) can be interpreted as the "accuracy for events + accuracy for non-events − 1". Hence, it does not depend on climatological frequency:

$$HK = \frac{H}{H + M} + \frac{F}{F + R} = POD - POFD.$$

Appendix A.2. Continuous Scores for the Rain Rate Estimation

In this section, the continuous scores employed in this study are explained and additional information on them can be found in the book of Wilks [73,74]. In the following, y_i denotes individual predictions and o_i the corresponding observations. The sample mean is calculated as follows for the predictions: $\bar{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$ and analogously for the observations $\bar{o}$. The sample variance of the predictions is expressed as: $s_y^2 = \frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y})^2$ and analogously for the observations s_o^2. Based on these variables, all the employed continuous scores can be computed.

The Mean Error (ME, also called Bias) measures the average prediction bias and denotes the average error without indicating the magnitude of the errors or how well the predictions and the observations correspond:

$$ME = \bar{y} - \bar{o}.$$

The Root Mean Squared Error (RMSE) represents the average magnitude of the prediction error weighted by the square of the error. It is sensitive to outliers because large errors are penalized heavily:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - o_i)^2}.$$

The Reduction of Variance skill score (RV, also called Nash–Sutcliffe Efficiency) measures the mean squared error of the predictions with reference to the mean squared error of a climatological prediction:

$$RV = 1 - \frac{MSE}{MSE_{clim}} = 1 - \frac{\frac{1}{N} \sum_{i=1}^{N} (y_i - o_i)^2}{s_o^2},$$

where MSE_{clim} is the climatological mean squared error represented by the observed sample variance.

The Pearson Correlation Coefficient (PCORR) depicts the sign and the strength of the linear relationship between the predicted and the observed variable with $|\text{PCORR}| = 1$ denoting a perfect linear relationship. PCORR is a measure of potential skill because PCORR^2 shows the proportion of the variance of the observations, which can be explained by a linear relationship with the predictions:

$$PCORR = \frac{\frac{1}{N} \sum_{i=1}^{N} (y_i - \bar{y}) \cdot (o_i - \bar{o})}{s_y \cdot s_o}.$$

The Spearman Rank Correlation (SCORR) measures the sign and the strength of a monotonic relationship, which can also be nonlinear. It is computed in the same way as PCORR but uses the ranks of the data instead of the data themselves, which makes it more robust and resistant to outliers:

$$SCORR = \frac{\frac{1}{N}\sum_{i=1}^{N}(y_i^r - \bar{y^r}) \cdot (o_i^r - \bar{o^r})}{s_y^r \cdot s_o^r},$$

where y_i^r is the rank of the ith prediction and o_i^r of the ith observations.

References

1. Schamm, K.; Ziese, M.; Becker, A.; Finger, P.; Meyer-Christoffer, A.; Schneider, U.; Schröder, M.; Stender, P. Global Gridded Precipitation over Land: A Description of the New GPCC First Guess Daily Product. *Earth Syst. Sci. Data* **2014**, *6*, 49–60, doi:10.5194/essd-6-49-2014.
2. Isotta, F.A.; Frei, C.; Weilguni, V.; Perčec Tadić, M.; Lassègues, P.; Rudolf, B.; Pavan, V.; Cacciamani, C.; Antolini, G.; Ratto, S.M.; et al. The Climate of Daily Precipitation in the Alps: Development and Analysis of a High-Resolution Grid Dataset from Pan-Alpine Rain-Gauge Data. *Int. J. Climatol.* **2014**, *34*, 1657–1675, doi:10.1002/joc.3794.
3. Crum, T.D.; Alberty, R.L. The WSR-88D and the WSR-88D Operational Support Facility. *Bull. Am. Meteorol. Soc.* **1993**, *74*, 1669–1688, doi:10.1175/1520-0477(1993)074<1669:TWATWO>2.0.CO;2.
4. Huuskonen, A.; Saltikoff, E.; Holleman, I. The Operational Weather Radar Network in Europe. *Bull. Am. Meteorol. Soc.* **2014**, *95*, 897–907, doi:10.1175/BAMS-D-12-00216.
5. Iguchi, T.; Kozu, T.; Meneghini, R.; Awaka, J.; Okamoto, K. Rain-Profiling Algorithm for the TRMM Precipitation Radar. *J. Appl. Meteorol.* **2000**, *39*, 2038–2052, doi:10.1175/1520-0450(2001)040<2038:RPAFTT>2.0.CO;2.
6. Joyce, R.J.; Janowiak, J.E.; Arkin, P.A.; Xie, P. CMORPH: A Method that Produces Global Precipitation Estimates from Passive Microwave and Infrared Data at High Spatial and Temporal Resolution. *J. Hydrometeorol.* **2004**, *5*, 487–503, doi:10.1175/1525-7541(2004)005<0487:CAMTPG>2.0.CO;2.
7. 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. The TRMM Multisatellite Precipitation Analysis (TMPA): Quasi-Global, Multiyear, Combined-Sensor Precipitation Estimates at Fine Scales. *J. Hydrometeorol.* **2007**, *8*, 38–55, doi:10.1175/JHM560.1.
8. Mugnai, A.; Casella, D.; Cattani, E.; Dietrich, S.; Laviola, S.; Levizzani, V.; Panegrossi, G.; Petracca, M.; Sanò, P.; Di Paola, F.; et al. Precipitation products from the hydrology SAF. *Nat. Hazards Earth Syst. Sci.* **2013**, *13*, 1959–1981, doi:10.5194/nhess-13-1959-2013.
9. Goudenhoofdt, E.; Delobbe, L. Evaluation of radar-gauge merging methods for quantitative precipitation estimates. *Hydrol. Earth Syst. Sci.* **2009**, *13*, 195–203, doi:10.5194/hess-13-195-2009.
10. Sideris, I.V.; Gabella, M.; Erdin, R.; Germann, U. Real-time radar—Rain-gauge merging using spatio-temporal co-kriging with external drift in the alpine terrain of Switzerland. *Q. J. R. Meteorol. Soc.* **2014**, *140*, 1097–1111, doi:10.1002/qj.2188.
11. Gourley, J.J.; Maddox, R.A.; Howard, K.W.; Burgess, D.W. An Exploratory Multisensor Technique for Quantitative Estimation of Stratiform Rainfall. *J. Hydrometeorol.* **2002**, *3*, 166–180, doi:10.1175/1525-7541(2002)003<0166:AEMTFQ>2.0.CO;2.
12. Adler, R.F.; Huffman, G.J.; Chang, A.; Ferraro, R.; Xie, P.; Janowiak, J.; Rudolf, B.; Schneider, U.; Curtis, S.; Bolvin, D.; et al. The Version-2 Global Precipitation Climatology Project (GPCP) Monthly Precipitation Analysis (1979–Present). *J. Hydrometeorol.* **2003**, *4*, 1147–1167, doi:10.1175/1525-7541(2003)004<1147:TVGPCP>2.0.CO;2.
13. Manz, B.; Buytaert, W.; Zulkafli, Z.; Lavado, W.; Willems, B.; Robles, L.A.; Rodríguez-Sánchez, J.P. High-resolution satellite-gauge merged precipitation climatologies of the tropical andes. *J. Geophys. Res. Atmos.* **2006**, *121*, 1190–1207, doi:10.1002/2015JD023788.
14. Haddad, Z.S.; Smith, E.A.; Kummerow, C.D.; Iguchi, T.; Farrar, M.R.; Durden, S.L.; Alves, M.; Olson, W.S. The TRMM 'Day-1' Radar/Radiometer Combined Rain-Profiling Algorithm. *J. Meteorol. Soc. Jpn.* **1997**, *75*, 799–809, doi:10.2151/jmsj1965.75.4_799.

15. Hou, A.Y.; Kakar, R.K.; Neeck, S.; Azarbarzin, A.A.; Kummerow, C.D.; Kojima, M.; Oki, R.; Nakamura, K.; Iguchi, T. The Global Precipitation Measurement Mission. *Bull. Am. Meteorol. Soc.* **2014**, *95*, 701–722, doi:10.1175/BAMS-D-13-00164.1.
16. Nerini, D.; Zulkafli, Z.; Wang, L.; Onof, C.; Buytaert, W.; Lavado-Casimiro, W.; Guyot, J. A Comparative Analysis of TRMM–Rain Gauge Data Merging Techniques at the Daily Time Scale for Distributed Rainfall-Runoff Modeling Applications. *J. Hydrometeorol.* **2015**, *16*, 2153–2168, doi:10.1175/JHM-D-14-0197.1.
17. Kitchen, M.; Blackall, R.M. Representativeness errors in comparisons between radar and gauge measurements of rainfall. *J. Hydrol.* **1992**, *134*, 13–33, doi:10.1016/0022-1694(92)90026-R.
18. Villarini, G.; Mandapaka, P.V.; Krajewski, W.F.; Moore, R.J. Rainfall and sampling uncertainties: A rain gauge perspective. *J. Geophys. Res. Atmos.* **2008**, *113*, D11102, doi:10.1029/2007JD009214.
19. Chen, M.; Shi, W.; Xie, P.; Silva, V.B. S.; Kousky, V.E.; Higgins, R.W.; Janowiak, J.E. Assessing objective techniques for gauge-based analyses of global daily precipitation. *J. Geophys. Res. Atmos.* **2008**, *113*, D04110, doi:10.1029/2007JD009132.
20. Nešpor, V.; Sevruk, B. Estimation of Wind-Induced Error of Rainfall Gauge Measurements Using a Numerical Simulation. *J. Atmos. Ocean. Technol.* **1999**, *16*, 450–464, doi:10.1175/1520-0426(1999)016<0450:EOWIEO>2.0.CO;2.
21. Kochendorfer, J.; Nitu, R.; Wolff, M.; Mekis, E.; Rasmussen, R.; Baker, B.; Earle, M.E.; Reverdin, A.; Wong, K.; Smith, C.D.; et al. Analysis of single-Alter-shielded and unshielded measurements of mixed and solid precipitation from WMO-SPICE. *Hydrol. Earth Syst. Sci.* **2017**, *21*, 3525–3542, doi:10.5194/hess-21-3525-2017.
22. Villarini, G.; Krajewski, W.F. Review of the Different Sources of Uncertainty in Single Polarization Radar-Based Estimates of Rainfall. *Surv. Geophys.* **2010**, *31*, 107–129, doi:10.1007/s10712-009-9079-x.
23. Berne, A.; Krajewski, W.F. Radar for hydrology: Unfulfilled promise or unrecognized potential? *Adv. Water Resour.* **2013**, *51*, 357–366, doi:10.1016/j.advwatres.2012.05.005.
24. Vasiloff, S.V.; Seo, D.; Howard, K.W.; Zhang, J.; Kitzmiller, D.H.; Mullusky, M.G.; Krajewski, W.F.; Brandes, E.A.; Rabin, R.M.; Berkowitz, D.S.; et al. Improving QPE and very short term QPF: An initiative for a community-wide integrated approach. *Bull. Am. Meteorol. Soc.* **2007**, *88*, 1899–1911, doi:10.1175/BAMS-88-12-1899.
25. de Coning, E.; Poolman, E. South African Weather Service operational satellite based precipitation estimation technique: Applications and improvements. *Hydrol. Earth Syst. Sci.* **2011**, *15*, 1131–1145, doi:10.5194/hess-15-1131-2011.
26. de Coning, E. Optimizing Satellite-Based Precipitation Estimation for Nowcasting of Rainfall and Flash Flood Events over the South African Domain. *Remote Sens.* **2013**, *5*, 5702–5724, doi:10.3390/rs5115702.
27. Scofield, R.A.; Kuligowski, R.J. Status and Outlook of Operational Satellite Precipitation Algorithms for Extreme-Precipitation Events. *Weather Forecast.* **2003**, *18*, 1037–1051, doi:10.1175/1520-0434(2003)018<1037:SAOOOS>2.0.CO;2.
28. Anagnostou, E.N. Overview of overland satellite rainfall estimation for hydro-meteorological applications. *Surv. Geophys.* **2004**, *25*, 511–537, doi:10.1007/s10712-004-5724-6.
29. Ebert, E.E.; Janowiak, J.E.; Kidd, C. Comparison of near-real-time precipitation estimates from satellite observations and numerical models. *Bull. Am. Meteorol. Soc.* **2007**, *88*, 47–64, doi:10.1175/BAMS-88-1-47.
30. Kidd, C.; Huffman, G. Global precipitation measurement. *Meteorol. Appl.* **2011**, *18*, 334–353, doi:10.1002/met.284.
31. Kidd, C.; Levizzani, V. Status of satellite precipitation retrievals. *Hydrol. Earth Syst. Sci.* **2011**, *15*, 1109–1116, doi:10.5194/hess-15-1109-2011.
32. Kidd, C.; Bauer, P.; Turk, J.; Huffman, G.J.; Joyce, R.; Hsu, K.; Braithwaite, D. Intercomparison of High-Resolution Precipitation Products over Northwest Europe. *J. Hydrometeorol.* **2012**, *13*, 67–83, doi:10.1175/JHM-D-11-042.1.
33. 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. Global-scale evaluation of 22 precipitation datasets using gauge observations and hydrological modeling. *Hydrol. Earth Syst. Sci.* **2017**, *21*, 6201–6217, doi:10.5194/hess-21-6201-2017.
34. Arkin, P.A.; Meisner, B.N. The Relationship between Large-Scale Convective Rainfall and Cold Cloud over the Western Hemisphere during 1982–1984. *Mon. Weather Rev.* **1987**, *115*, 51–74, doi:10.1175/1520-0493(1987)115<0051:TRBLSC>2.0.CO;2.

35. Ba, M.B.; Gruber, A. GOES Multispectral Rainfall Algorithm (GMSRA). *J. Appl. Meteorol.* **2001**, *40*, 1500–1514, doi:10.1175/1520-0450(2001)040<1500:GMRAG>2.0.CO;2.
36. Sorooshian, S.; Hsu, K.; Gao, X.; Gupta, H.V.; Imam, B.; Braithwaite, D. Evaluation of PERSIANN system satellite-based estimates of tropical rainfall. *Bull. Am. Meteorol. Soc.* **2000**, *81*, 2035–2046, doi:10.1175/1520-0477(2000)081<2035:EOPSSE>2.3.CO;2.
37. Ebert, E.E.; Manton, M.J.; Arkin, P.A.; Allam, R.J.; Holpin, G.E.; Gruber, A. Results from the GPCP Algorithm Intercomparison Programme. *Bull. Am. Meteorol. Soc.* **1996**, *77*, 2875–2887, doi:10.1175/1520-0477(1996)077<2875:RFTGAI>2.0.CO;2.
38. Vicente, G.A.; Scofield, R.A.; Menzel, W.P. The operational GOES infrared rainfall estimation technique. *Bull. Am. Meteorol. Soc.* **1998**, *79*, 1883–1898, doi:10.1175/1520-0477(1998)079<1883:TOGIRE>2.0.CO;2.
39. Adler, R.F.; Negri, A.J.; Keehn, P.R.; Hakkarinen, I.M. Estimation of monthly rainfall over Japan and surrounding waters from a combination of low-orbit microwave and geosynchronous IR data. *J. Appl. Meteorol.* **1993**, *32*, 335–356, doi:10.1175/1520-0450(1993)032<0335:EOMROJ>2.0.CO;2.
40. Meyer, H.; Kühnlein, M.; Appelhans, T.; Nauss, T. Comparison of four machine learning algorithms for their applicability in satellite-based optical rainfall retrievals. *Atmos. Res.* **2016**, *169*, 424–433, doi:10.1016/j.atmosres.2015.09.021.
41. Thies, B.; Nauss, T.; Bendix, J. Discriminating raining from non-raining cloud areas at mid-latitudes using meteosat second generation SEVIRI night-time data. *Meteorol. Appl.* **2008**, *15*, 219–230, doi:10.1002/met.56.
42. Thies, B.; Nauss, T.; Bendix, J. Discriminating raining from non-raining clouds at mid-latitudes using meteosat second generation daytime data. *Atmos. Chem. Phys.* **2008**, *8*, 2341–2349, doi:10.5194/acp-8-2341-2008.
43. Roebeling, R.A.; Holleman, I. SEVIRI rainfall retrieval and validation using weather radar observations. *J. Geophys. Res. Atmos.* **2009**, *114*, D21202, doi:10.1029/2009JD012102.
44. Kühnlein, M.; Thies, B.; Nauss, T.; Bendix, J. Rainfall-Rate Assignment Using MSG SEVIRI Data—A Promising Approach to Spaceborne Rainfall-Rate Retrieval for Midlatitudes. *J. Appl. Meteorol. Climatol.* **2010**, *49*, 1477–1495, doi:10.1175/2010JAMC2284.1.
45. Capacci, D.; Conway, B.J. Delineation of precipitation areas from MODIS visible and infrared imagery with artificial neural networks. *Meteorol. Appl.* **2005**, *12*, 291–305, doi:10.1017/S1350482705001787.
46. Behrangi, A.; Hsu, K.; Imam, B.; Sorooshian, S.; Kuligowski, R.J. Evaluating the Utility of Multispectral Information in Delineating the Areal Extent of Precipitation. *J. Hydrometeorol.* **2009**, *10*, 684–700, doi:10.1175/2009JHM1077.1.
47. Kühnlein, M.; Appelhans, T.; Thies, B.; Nauss, T. Improving the accuracy of rainfall rates from optical satellite sensors with machine learning—A random forests-based approach applied to MSG SEVIRI. *Remote Sens. Environ.* **2014**, *121*, 129–143, doi:10.1016/j.rse.2013.10.026.
48. Hong, Y.; Hsu, K.; Sorooshian, S.; Gao, X. Precipitation Estimation from Remotely Sensed Imagery Using an Artificial Neural Network Cloud Classification System. *J. Appl. Meteorol.* **2004**, *43*, 1834–1853, doi:10.1175/JAM2173.1.
49. Tao, Y.; Gao, X.; Ihler, A.; Sorooshian, S.; Hsu, K. Precipitation Identification with Bispectral Satellite Information Using Deep Learning Approaches. *J. Hydrometeorol.* **2017**, *18*, 1271–1283, doi:10.1175/JHM-D-16-0176.1.
50. Tapiador, F.J.; Kidd, C.; Levizzani, V.; Marzano, F. A Neural Networks-Based Fusion Technique to Estimate Half-Hourly Rainfall Estimates at 0.1° Resolution from Satellite Passive Microwave and Infrared Data. *J. Appl. Meteorol.* **2004**, *43*, 576–594, doi:10.1175/1520-0450(2004)043<0576:ANNFTT>2.0.CO;2.
51. Kacimi, S.; Viltard, N.; Kirstetter, P.E. A new methodology for rain identification from passive microwave data in the Tropics using neural networks. *Q. J. R. Meteorol. Soc.* **2013**, *139*, 912–922, doi:10.1002/qj.2114.
52. Tapiador, F.J.; Kidd, C.; Hsu, K.; Marzano, F. Neural networks in satellite rainfall estimation. *Meteorol. Appl.* **2004**, *11*, 83–91, doi:10.1017/S1350482704001173.
53. Islam, T.; Rico-Ramirez, M.; Srivastava, P.K.; Dai, Q. Non-parametric rain/no rain screening method for satellite-borne passive microwave radiometers at 19–85 GHz channels with the Random Forests algorithm. *Int. J. Remote Sens.* **2014**, *35*, 3254–3267, doi:10.1080/01431161.2014.903444.
54. Kühnlein, M.; Appelhans, T.; Thies, B.; Nauss, T. Precipitation estimates from MSG SEVIRI daytime, nighttime, and twilight data with random forests. *J. Appl. Meteorol. Climatol.* **2014**, *53*, 2457–2480, doi:10.1175/JAMC-D-14-0082.1.

55. Schmetz, J.; Pili, P.; Tjemkes, S.; Just, D.; Kerkmann, J.; Rota, S.; Ratier, A. An introduction to Meteosat second generation (MSG). *Bull. Am. Meteorol. Soc.* **2002**, *83*, 977–992, doi:10.1175/1520-0477(2002)083<0977:AITMSG>2.3.CO;2.
56. EUMETSAT NWC-SAF: Support to Nowcasting and Very Short Range Forecasting. Available online: nwcsaf.org (accessed on 16 May 2018).
57. Derrien, M.; Le Gléau, H.; Fernandez, P. *Algorithm Theoretical Basis Document for Cloud Products (CMa-PGE01 v3. 2, CT-PGE02 v2. 2 & CTTHPGE03 v2. 2)*; Technical Report; 2013; pp. 1–87. Available online: SAF/NWC/CDOP2/MFL/SCI/ATBD/01 (accessed on 16 May 2018).
58. Raspaud, M.; Dybbroe, A.; Devasthale, A.; Lahtinen, P.; Rasmussen, L.Ø.; Hoese, D.; Nielsen, E.; Leppelt, T.; Maul, A.; Hamann, U.; et al. PyTroll: An open source, community driven Python framework to process Earth Observation satellite data. *Bull. Am. Meteorol. Soc.* **2018**, doi:10.1175/BAMS-D-17-0277.1.
59. EUMETNET OPERA. Available online: eumetnet.eu/opera (accessed on 16 May 2018).
60. Gabella, M.; Joss, J.; Perona, G. Optimizing quantitative precipitation estimates using a noncoherent and a coherent radar operating on the same area. *J. Geophys. Res. Atmos.* **2000**, *105*, 2237–2245, doi:10.1029/1999JD900420.
61. Gabella, M.; Joss, J.; Perona, G.; Michaelides, S. Range adjustment for ground-based radar, derived with the spaceborne TRMM precipitation radar. *IEEE Trans. Geosci. Remote Sens.* **2006**, *44*, 126–133, doi:10.1109/TGRS.2005.858436.
62. Germann, U.; Joss, J. Operational Measurements of Precipitation in Mountainous Terrain. In *Weather Radar*; Meischner, R., Ed.; Springer: Berlin, Germany, 2004; pp. 52–77, ISBN 978-3-662-05202-0.
63. Koistinen, J.; Michelson, D.B.; Hohti, H.; Peura, M. Operational Measurement of Precipitation in Cold Climates. In *Weather Radar*; Meischner, R., Ed.; Springer: Berlin, Germany, 2004; pp. 78–114, ISBN 978-3-662-05202-0.
64. Zhang, J.; Qi, Y.; Kingsmill, D.; Howard, K. Radar-Based Quantitative Precipitation Estimation for the Cool Season in Complex Terrain: Case Studies from the NOAA Hydrometeorology Testbed. *J. Hydrometeorol.* **2012**, *13*, 1836–1854, doi: 10.1175/JHM-D-11-0145.1.
65. EUMETSAT H-SAF: Support to Operational Hydrology and Water Management. Available online: hsaf.meteoam.it (accessed on 16 May 2018).
66. *Product User Manual for product H03A – PR-OBS-3A (Version 1.2): Precipitation Rate at Ground by GEO/IR Supported by LEO/MW*; Technical Report; 2015; pp. 1–19. Available online: SAF/HSAF/PUM-03A (accessed on 16 May 2018).
67. Dormann, C.F.; Elith, J.; Bacher, S.; Buchmann, C.; Carl, G.; Carré, G. Collinearity: A review of methods to deal with it and a simulation study evaluating their performance. *Ecography* **2013**, *36*, 27–46, doi:10.1111/j.1600-0587.2012.07348.x.
68. Scikit-Learn: Machine Learning in Python. Available online: scikit-learn.org (accessed on 16 May 2018).
69. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
70. Meyer, H.; Kühnlein, M.; Reudenbach, C.; Nauss, T. Revealing the potential of spectral and textural predictor variables in a neural network-based rainfall retrieval technique. *Remote Sens. Lett.* **2017**, *8*, 647–656, doi:10.1080/2150704X.2017.1312026.
71. Schleiss, M.; Chamoun, S.; Berne, A. Nonstationarity in Intermittent Rainfall: The "Dry Drift". *J. Hydrometeorol.* **2014**, *15*, 1189–1204, doi:10.1175/JHM-D-13-095.1.
72. Hogan, R.J.; Mason, I.B. Deterministic forecasts of binary events. In *Forecast Verification: A Practitioner's Guide in Atmospheric Science*; Jolliffe, I.T., Stephenson, D.B., Eds.; John Wiley & Sons: Chichester, UK, 2012; pp. 31–59, ISBN 978-0470660713.
73. Wilks, D.S. Empirical Distributions and Exploratory Data Analysis. *Int. Geophys.* **2011**, *100*, 23–70, doi:10.1016/B978-0-12-385022-5.00003-8.
74. Wilks, D.S. Forecast Verification. *Int. Geophys.* **2011**, *100*, 301–394, doi:10.1016/B978-0-12-385022-5.00008-7.