


Solar Radiation Nowcasting Using a Markov Chain Multi-Model Approach

Xinyuan Hou ^{1,2*} , Kyriakoula Papachristopoulou ^{1,3,4}, Yves-Marie Saint-Drenan ⁵ and Stelios Kazadzis ¹

¹ Physikalisch-Meteorologisches Observatorium Davos/World Radiation Center (PMOD/WRC), 7260 Davos Dorf, Switzerland

² Department of Physics, ETH Zürich, 8093 Zürich, Switzerland

³ Laboratory of Climatology and Atmospheric Environment, Sector of Geography and Climatology, Department of Geology and Environment, National and Kapodistrian University of Athens (LACAE/NKUA), Athens 15772, Greece

⁴ Institute for Astronomy, Astrophysics, Space Applications and Remote Sensing, National Observatory of Athens (IAASARS/NOA), Athens 11810, Greece

⁵ O.I.E. Centre Observation, Impacts, Energy, MINES ParisTech, PSL Research University, 06904 Sophia Antipolis, France

* Correspondence: xinyuan.hou@pmodwrc.ch

Abstract: Solar resources find increasing application in the recent years and the demand will continue to grow as the society directs to a more renewable development path. However, the required high-frequency solar irradiance data is not yet readily available everywhere and there has been endeavors to its forecasting to facilitate grid integration, such as in the photovoltaics power planning. The objective of this study is to develop a hybrid approach to improve the accuracy of solar nowcasting with the lead time of up to one hour. The proposed method utilizes the irradiance data from the Copernicus Atmospheric Monitoring Service (CAMS) for four European cities with various cloud conditions. The approach effectively improves the prediction accuracy in all four cities. In the prediction of global horizontal irradiance for Berlin, the reduction of mean daily error during a month amounts to 2.5 Wh m^{-2} and the monthly relative improvement reaches nearly 5% compared with the traditional persistence method. In the other three cities, accuracy improvement can also be observed. Furthermore, since the proposed approach only requires solar radiation data as model input, which can be conveniently obtained from CAMS, it possesses the potential to be upscaled at regional level in response to the needs of the pan-EU energy transition.

Keywords: solar radiation nowcasting; solar energy prediction; Markov chain models

Citation: Hou, X.; Papachristopoulou, K.; Saint-Drenan, Y.-M.; Kazadzis, S. Solar Radiation Nowcasting Using a Markov Chain Multi-Model Approach. *Energies* **2022**, *1*, 0. <https://doi.org/>

Received:

Accepted:

Published:

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Submitted to *Energies* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Global horizontal irradiance (GHI) is defined as the downwelling solar irradiance observed at ground level on horizontal surfaces and integrated over the whole shortwave spectrum. It is the sum of the direct and diffuse irradiance, which come from the direction of the sun and from the rest of the sky, respectively [1]. Global solar irradiation data are crucial for renewable energy applications including the planning, monitoring and forecasting of the photovoltaics (PV) systems. However, irradiation data are not readily available everywhere, especially when meteorological stations lack in isolated sites [2]. Ngoko et al. [3] also noted that historical solar radiation data taken at high sampling frequencies are unavailable for many locations, thereby limiting some just-in-time applications to specific sites where such data might not exist.

Like for the majority of the meteorological variables, a wide spectrum of forecasting techniques finds current application in solar radiation forecasting. One group of solar forecasting methods consists of statistical solar forecasts based on irradiance time series [4] or machine learning methods such as artificial neural networks (ANN) [5]. Another group includes physical methods such as irradiance forecasting with cloud motion vectors (CMV) [6], or with numerical weather prediction. Both groups employ also extensively

post-processing methods such as the use of model output statistics [5], quantile forecasts [7] and ensemble prediction system [8]. Statistical post-processing techniques learn error pattern by comparing forecasts and observations to reduce the error in the final prediction [9].

Mainly driven by synoptic and local weather patterns, surface solar radiation is highly variable. While clear-sky irradiance is strongly influenced by atmospheric composition including aerosol and water vapor, all-sky irradiance changes with clouds, resulting in a high fluctuation of solar radiation. Considering the variability of the cloudiness and solar irradiance, probabilistic techniques, such as Markov chain based methods, are widely used in solar forecasting. For example, a homogeneous recurrent Markov process with discrete states aided in predicting sunshine and cloud cover for Payern, Switzerland and Perth, Australia [10]. Hocaoglu et al. [11] combined the Mycielski algorithm with a Markov chain model to forecast hourly solar radiation from historical record for two regions in Turkey. A Markov chain probability distribution mixture model was employed for the forecast of clear-sky index in Hawaii and Norrköping [12,13].

Besides solar radiation forecasting, synthetic generation of solar radiation time series has also been of interest and can be found in some previous studies: Poggi et al. [14] studied stochastic property of generated hourly total solar radiation in Corsica, France using a Markov model following the shifted negative binomial distribution. Mellit et al. [2] used an ANN and Markov transition matrices for the generation of daily global solar radiation. Ngoko et al. [3] generated high temporal resolution (1-minute) solar radiation using Markov models with input of hourly sea level pressure, wind speed, cloud based height and cloud cover for two locations in Japan. Bright et al. [15] set up a Markov chain with hourly weather observation data of cloud cover in Leeds, UK to generate minutely irradiance time series. Shepero et al. [16] employed hidden Markov models with Gaussian observation distributions to generate the clear-sky index time series for two locations with different climate conditions (Hawaii and Norrköping). Urrego-Ortiz et al. [17] proposed a Markov chain model for day-ahead forecasting of hourly GHI in Medellín, Colombia.

Nowcasting systems correspond to the forecast of the short term evolution of the weather of up to 6 hours ahead [11]. It is of importance to provide information about solar radiation availability in the next hour(s) to facilitate solar energy management by plant operators and grid operators. In this study, we propose a hybrid approach utilizing several forecast techniques, including but not limited to the persistence ensemble and the Markov chain, in order to increase the accuracy of solar radiation nowcasting. Combining different models in an entity is expected to leverage on the strengths of the best methods while reducing their weaknesses.

2. Data

We use radiation data provided by the Copernicus Atmospheric Monitoring Service (CAMS) radiation service for its wide geographical coverage from -66° to 66° in both latitudes and longitudes [18]. The CAMS radiation data are outputs using the Heliostat-4 method [19] and the fast clear-sky model McClear [1,20]. McClear utilizes aerosol properties and total column content of water vapor and ozone obtained from the CAMS model. Besides the solar zenith angle and the ground albedo, six other parameters describing the optical state of the atmosphere are used as input to the McClear look-up tables to estimate downwelling shortwave radiation in cloudless conditions.

Previous studies have identified that the interactions between clouds and radiation contribute to the stochastic variations of GHI (e.g., [17]). Here, this study considers the Cloud modification factor (CMF), defined as the ratio of GHI in all-sky condition to that in clear-sky conditions (GHI_{CS}) keeping the same atmospheric composition but assuming the absence of clouds:

$$CMF = \frac{GHI}{GHI_{CS}} \quad (1)$$

This measure is similar to the clearness index, a ratio of the solar radiation recorded on the earth's surface to the extraterrestrial solar radiation. If CMF approaches 0, the sky is cloudy, while a CMF value of 1 implies that it is clear sky without clouds.

The data from CAMS are updated until two days prior to the instantaneous query time. We take the time series with an interval of 15 minutes, extending from 2004-02-01 to 2020-12-31. Thanks to the wide coverage of CAMS data, we choose four cities of different cloud conditions in Europe, listed in Table 1:

Table 1. Four chosen cities with their coordinates, long-term cloud modification factor (CMF, mean \pm standard deviation from 2004 to 2019) and mean absolute change of CMF from the previous one time step to the present.

City	Coordinates	long-term CMF	$ \Delta\text{CMF} $
Athens	37.98° N, 23.72° E	0.787 ± 0.245	0.049
Bucharest	44.44° N, 26.08° E	0.693 ± 0.278	0.054
Berlin	52.52° N, 13.37° E	0.656 ± 0.267	0.067
Helsinki	60.19° N, 24.93° E	0.625 ± 0.267	0.063

which are in descending order of long-term average CMF. The last column of Table 1 presents the mean absolute change of CMF from the previous one time step to the present as a measure of cloudiness variability. Athens has comparatively low variability in cloudiness, whereas Berlin has the highest variability among the four cities and thus is the most difficult to predict. Figure 1 shows the daily mean insolation (GHI integrated during a day) for each month and monthly mean CMF for Berlin in 2020, as an example.

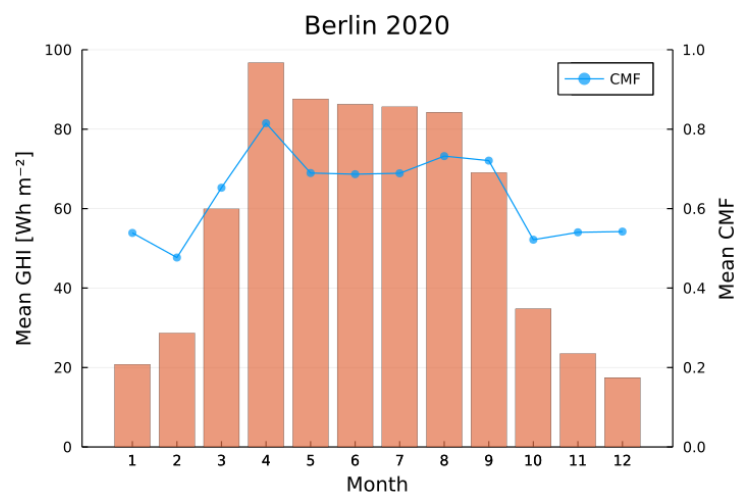


Figure 1. Daily mean global horizontal irradiation (GHI) for each month (orange bars, left axis) and monthly mean cloud modification factor (CMF, blue curves, right axis) for Berlin in 2020. The results for the 3 other cities can be found in Figure A1 in the Appendix.

We divide the CMF time series into the period from 2004 to 2018 for fitting the transition matrix from historical occurrence (training set), the year 2019 for evaluating the error metrics of the prediction methods (validation set), and the year 2020 for testing the performance of different prediction approaches (test set), including the implementation of the hybrid approach (see Section 3.5).

3. Model and Methods

3.1. Quantitative Metrics

In this study, we adopt several commonly used metrics for the selection of model parameters as well as accuracy assessment. Firstly, the error of one data pair is defined as the difference between the predicted (\hat{X}_i) and the actual value (X_i):

$$\epsilon_i = \hat{X}_i - X_i \quad (2)$$

Standard deviation is given by:

$$\sigma(\epsilon) = \sqrt{\frac{1}{N} \sum_{i=1}^N (\epsilon_i - \bar{\epsilon})^2} \quad (3)$$

where N is the number of data pairs in the time series and $\bar{\epsilon}$ denotes the mean error.

Pearson Correlation coefficient is defined as:

$$r = \frac{\sum_{i=1}^N (\hat{X}_i - \bar{\hat{X}})(X_i - \bar{X})}{\sigma(\hat{X})\sigma(X)} \quad (4)$$

where $\bar{\hat{X}}$ and \bar{X} represent the mean value of the predicted and the actual time series, respectively.

Mean bias error often simply referred to as *bias*, is given by

$$MBE = \frac{1}{N} \sum_{i=1}^N \epsilon_i \quad (5)$$

Mean absolute error is written:

$$MAE = \frac{1}{N} \sum_{i=1}^N |\epsilon_i| \quad (6)$$

Root mean square error is defined as:

$$RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^N \epsilon_i^2} \quad (7)$$

3.2. Markov Chain

A discrete-time Markov chain describes the transition process of a discrete random variable, where the probability distribution of the following state of the variable depends on its previous states. Markov chain based methods are frequently used for forecast when the lead time is approximately equal to the temporal resolution of the model, i.e., a few time steps into the future. Solar radiation derived CMF possesses Markov property, which indicates that the evolution of the Markov process in the future depends only on the present state or plus a few past states, depending on the order of the Markov chain. This property permits to predict solar radiation via a Markov chain (MC) model.

The order of the Markov process indicates the number of previous observations on which the next state of the variable statistically depends [17]. In an i -th order Markov chain model, the probability that the process will be in a particular state at time t depends not only on its state at time $t-1$ but also on the states at times $t-2$; $t-3$; ...; $t-i$. Therefore, for instance, a second order Markov model is only an approximation of the process generating the observed data which is actually infinitely complex [3]. To check the stationarity of a Markov process, one can divide the sequence of the events into a few sub-intervals and compare the transition probability matrix of each sub-interval [14]. Thus, one possibility is to consider

seasonality when constructing the MC, such as computing four transition matrices for each of the four seasons throughout the year. However, to reduce the complexity and retain the generality of the analysis, we do not construct the Markov chain on a seasonal basis in the present study.

3.2.1. State Classification

According to the definition of CMF in Section 2, it can take any value between 0 and 1. To apply CMF in the MC model, the continuous variable firstly need to be discretized into several classes. The next step involves the setting of the number of class n , which reflects a trade-off between the accuracy and complexity of the second order MC model. Figure 2 shows the standard deviation of the MC prediction bias for Berlin in 2019 for one to four time steps ahead in the year 2019 using different number of CMF class from 5 to 40 classes in an interval of 5. While 40 classes deliver even lower biases, we opt for 30 classes in order not to overfit the training set in case that the classification generalizes unsatisfactorily, for example too specific for the historical record to be applied to a later time.

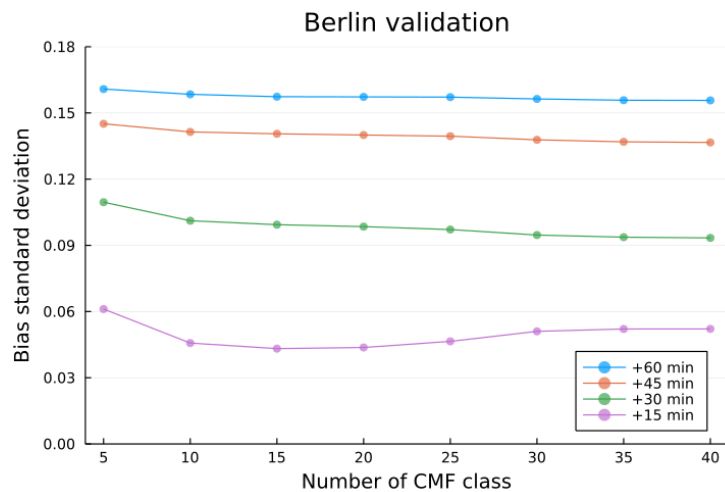


Figure 2. Bias standard deviation of Markov chain (MC) model with different number of classes for the validation set (2019) in Berlin.

Therefore, we categorize the training CMF values into 30 classes, each having the same amount of data points. Consequently, the ranges of CMF classes are not equal, but in larger (smaller) intervals for smaller (higher) CMF values towards 0 (towards 1). We compute the mean CMF value within each class and these 30 representative values are used for the prediction in combination with the transition matrix. Figure 3 displays the distributions of each CMF class for the three data sets of Berlin. The validation and test sets inherit the classification from the training set, thus do not have the equal number of data points in every class. The classes of very high CMF values (clear sky) are in the majority throughout the whole period.

3.2.2. Transition Matrix

The comparison of the standard deviation of prediction bias by 30-class MC models of first to third order is shown in Figure 4. Since the second order Markov chain, which considers two previous successive transition entries, has smaller standard deviations than the other two orders in every lead time step, we choose to use it. For the second order Markov chain, we initiate an $n^2 \times n$ matrix. The columns represent n possible states for the immediate time step ($n=30$ in this study), whereas the rows list all n^2 transition possibilities of states from the previous to the current time step, namely $1 \rightarrow 1, 1 \rightarrow 2, \dots, 1 \rightarrow 30, 2 \rightarrow 1, \dots$, and $30 \rightarrow 30$. We count the occurrence of each case to fill into the matrix. After converting the integer occurrences to frequencies within each row, we obtain the normalized transition matrix, indicating the frequency of the next possible state at a given transition from the

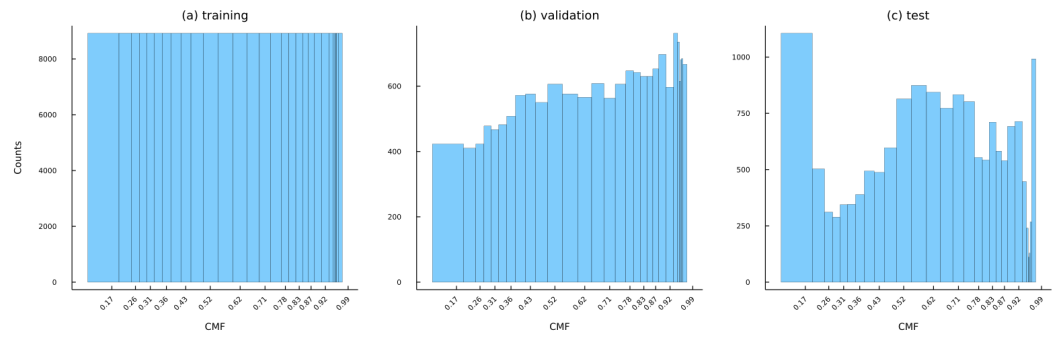


Figure 3. Distribution of CMF class of three data sets in Berlin: (a) training set for 2004-2018, (b) validation set for 2019 and (c) test set for 2020.

previous to the current time step. Figure 5 visualizes the prediction of the second order MC model for the validation set in Berlin, with the horizontal axis representing the CMF class from 2 time steps prior and the vertical axis 1 time step earlier. The color in each cell denotes the CMF value for the time step to be predicted. The pattern of vertical color bands on the matrix again supports the choice of the second order MC, since the CMF value at the current step closely follow the range from CMF 2 time steps earlier.

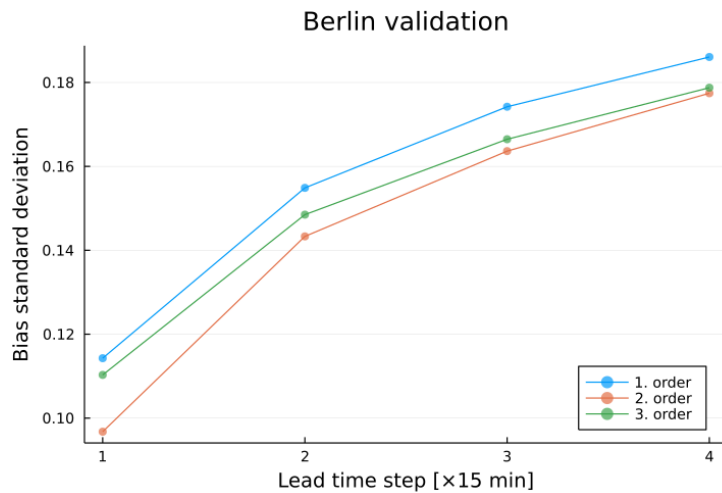


Figure 4. Bias standard deviation of MC model of different orders for the validation set (2019) in Berlin.

3.2.3. Temporal Extrapolation

Variant a From the validation or test set, we take 2 consecutive values as the observation and compute the value at the next step from the pre-calculated transition matrix. Rather than using a random sampling from a uniform distributed variable, as employed in some previous studies such as [17] and [14], we compute the prediction value \hat{X} by weighted mean at the specific row (transition) in the transition matrix, i.e., summing up the product of the average of each CMF class \overline{X}_i and the probability of each class p_i :

$$\hat{X} = \sum_{i=1}^N p_i \overline{X}_i \quad (8)$$

This predicted value is then set to be CMF at 1 to 4 time steps later. Figure 6 illustrates a diagram discerning two variants of temporal extrapolation during an one-hour time window from 9:00 to 10:00 in the MC model, in which the upper part specifies Variant a.

Variant b To predict the CMF values longer than 2 time steps ahead of the current time step, we take 2 consecutive values from the validation set as the observation, and

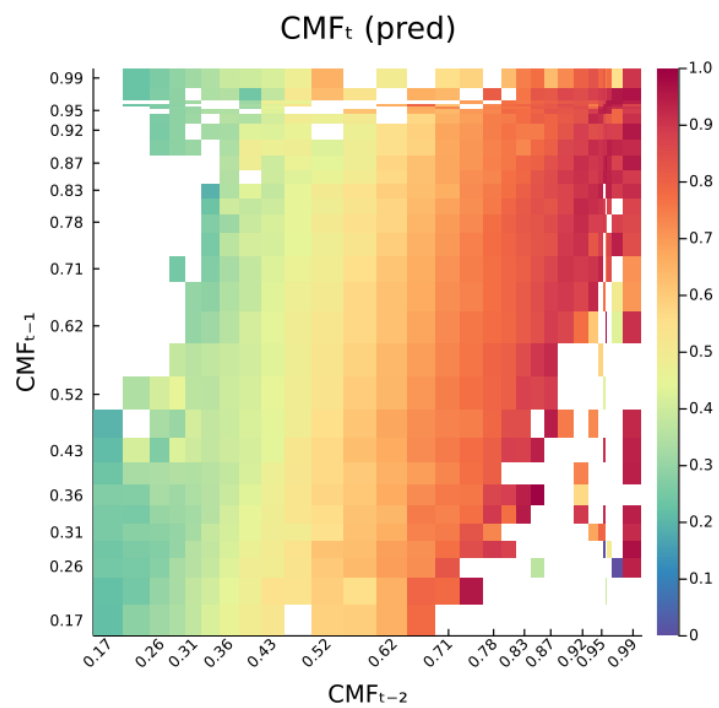


Figure 5. Prediction of the second order MC model for the validation set (2019) in Berlin.

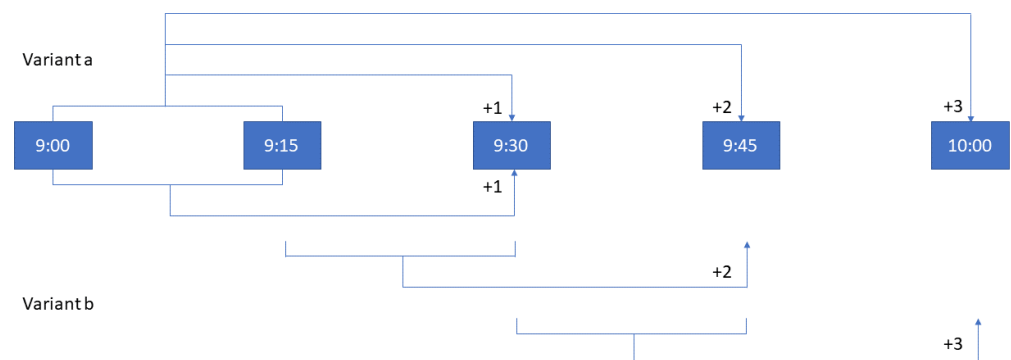


Figure 6. Diagram of two variants of temporal extrapolation during a one-hour time window from 9:00 to 10:00 in the MC model.

we predict the value for the next time step following Variant a. For the prediction of one further time step, the previous 2 observation points consist of the latter one from the record and the newly predicted one, and we apply the prediction following Variant a. Therefore, for one step ahead, the results of both variants are the same, but only Variant a would be visualized to avoid redundancy and/or overlapping in the figures. If the value at one further time step is needed, the new observation would be the 2 consecutive predictions. We repeat this procedure until reaching the prediction for 4 time steps ahead. In Variant b, during a time window of one hour, only the first two CMF values are taken from the record, the following inputs are either a combination of record and prediction, or solely predicted values (lower part in Figure 6).

3.3. Persistence Approach

The persistence approach assumes that the cloud state does not change from the previous to the current time step, therefore, it uses the CMF value at the previous time step as the prediction for the current time step. It is essentially a temporal shift of the original time series by a certain interval, in this case, from 1 to 4 time steps earlier to the present.

3.4. Neighbor Inference Approach

To trace the cloud movement above a region, one option is to look up the record with a temporal lag in its neighboring region. Using Berlin as a test bed, we evaluate the viability of inferring CMF values of the central grid cell from its neighbor cells at the previous time step. The neighbor cells are the central cell shifted to its eight directions by 0.1° longitude or/and latitude: to the west, northwest, north, northeast, east, southeast, south and southwest. We firstly exclude the time steps with CMF values higher than or equal to 0.95 (sunny with very few clouds) and compute the Pearson correlation of CMF values at the central cell with those at the neighbor cells to obtain eight correlation coefficients. Figure 7 exhibits the correlation coefficients of the CMF values in Berlin with CMF values from the previous time step (15 minutes prior) in the eight neighbor cells based on the time series from 2004 to 2019. The highest correlation is with the cell to the west, followed by the cells in the northwest and southwest. Moving towards the east, the correlation decreases and the minimum situates in the southeastern cell. This is as expected, since the general movement of clouds above Berlin is westerly flow. In the following, we evaluate the performance of using prior CMF values from the western cell in comparison to other prediction techniques.

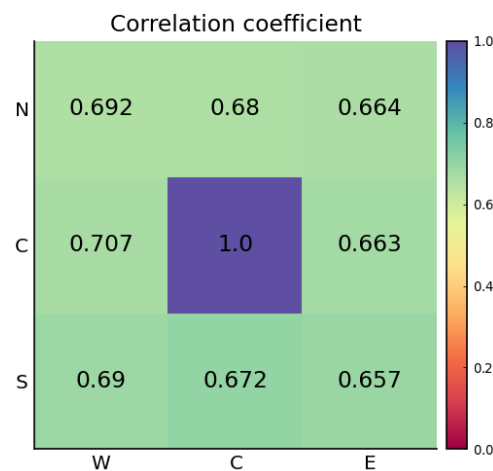


Figure 7. Correlation coefficients of CMF in Berlin with its eight neighboring grid cells.

3.5. Hybrid Approach

For the time series of 2019, we evaluate the accuracy of persistence method, neighbor inference, MC prediction Variant a and Variant b to develop hybrid methods based on MAE

or RMSE for all 30 CMF classes. MAE measures the mean difference between predicted and actual time series, whereas RMSE measures the bias and standard deviation of the difference between predicted and actual time series, and both have the same unit as the actual time series (dimensionless in the case of CMF). If the error by a certain approach is the smallest among the considered approaches, we adopt this approach for this specific CMF class. We apply this information on the time series of 2020 by choosing the optimal approach for the next time step based on the CMF class at the current time step. Table 2 gives an overview of the different approaches and their respective indices in Section 4.

Table 2. Overview of prediction methods with their indices and color code.

Method	Index	Color
Persistence	pers	violet
MC Variant a	mc_a	skyblue
MC Variant b	mc_b	pink
Neighbor inference	neib	olive
Hybrid by MAE	hyb_m	black
Hybrid by RMSE	hyb_r	red

3.6. GHI Prediction

For the computation of all-sky GHI, we multiply clear-sky GHI provided by CAMS by predicted CMF values following different approaches:

$$GHI = GHI_{CS} \cdot CMF \quad (9)$$

4. Results and Discussion

4.1. Results of CMF

In this section, we first present the results for the evaluation of CMF for Berlin and then for the other three cities. Table 3 lists Pearson correlation coefficients of predicted time series using different methods with the actual time series for 1 to 4 time steps ahead in 2020 for Berlin. Within each lead time step, the correlation coefficients between the prediction and the record is similar among different methods, ranging from around 0.9 for 1 time step ahead to approximately 0.7 for 4 time steps ahead. The MC model Variant a could outperform the persistence approach for the forecast horizon within half an hour. For a lead time of longer than 45 minutes, the neighbor inference wins an edge and has higher correlation with the actual time series than the persistence or MC predictions. Built upon these comparisons, the hybrid approaches outperform the other methods, where the one based on RMSE is slightly better than the one based on MAE, except for the lead time of one hour.

Table 3. Correlation coefficients of predicted with actual CMF time series for 4 lead time steps (+1 to +4) in 2020 for Berlin. The results for the 3 other cities can be found in Tables A1 to A3 in the Appendix.

Method	+1	+2	+3	+4
pers	0.9087	0.8142	0.7635	0.7275
mc_a	0.914	0.8204	0.7658	0.7254
mc_b	0.914	0.7956	0.72	0.6691
neib	0.8874	0.805	0.7653	0.7331
hyb_m	0.9143	0.8301	0.7946	0.7628
hyb_r	0.9146	0.8352	0.7975	0.7609

Next, we examine the agreement of the CMF values between different methods and the actual time series by the standard deviation of the prediction bias. Figure 8 shows the standard deviation of the bias in CMF from different approaches. In general, the standard deviation of the prediction bias for each approach increases with the lead time. For 1 time step ahead, standard deviations of the bias for hybrid approaches are very close to that for the persistence approach. As the lead time increases, the differences between them also increase. That is, the standard deviation of the bias by the persistence method increases more than that for hybrid approaches. When the lead time is 15 minutes, the standard deviation for hybrid approach based on RMSE is almost equal to that based on MAE, however, in later time steps, the ones based on RMSE have slightly lower standard deviations than the ones based on MAE.

In the first and second lead time step, the neighbor inference approach has the highest standard deviation of bias among the methods considered. For 3 and 4 time steps ahead, the standard deviation of bias becomes slightly smaller than that for the persistence method. The MC prediction Variant a has lower standard deviation of bias throughout than the persistence approach. For 1 time step ahead, the standard deviation from the MC Variant a is almost equal to that of hybrid approaches. On the other hand, the MC prediction Variant b has relatively large standard deviation for 15 to 60 minutes ahead of lead time. It only performs a little better than the neighboring grid cell for 2 time steps ahead, but renders the maximum standard deviation among all for the lead time of more than 30 minutes. For 1 hour ahead, the standard deviation for the MC prediction Variant b exceeds 0.2.

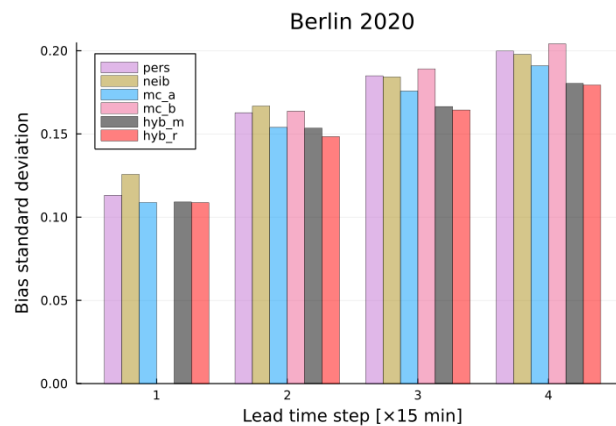


Figure 8. Standard deviation of CMF prediction bias for Berlin. The results for the 3 other cities can be found in Figure A2 in the Appendix.

Besides the comparison between the prediction and the actual time series, we also compare the errors of prediction between different methods by means of MAE and RMSE. Figure 9 shows the MAE and RMSE of CMF from different methods for four lead time steps in 2020 for Berlin from the lower left to upper right corner. For 15 minutes ahead, the errors of the hybrid methods are close to the persistence method. From 30 minutes ahead on, the errors of the hybrid methods are apparently smaller than the persistence method. The errors of the persistence method and the neighbor inference for longer than 45 minutes ahead are even larger than those of the hybrid methods for one hour ahead. While the MAEs of the MC prediction Variant a are for all lead time steps larger than the other methods except the MC prediction Variant b, its RMSE is on the comparable level with the other for one time step ahead and is evidently smaller than the persistence or neighboring approaches from 30 minutes ahead on. On the other hand, the MC prediction Variant b has no advantage over the other methods, regardless of MAE or RMSE across the three lead time steps applied.

The averaging of errors could mask the variation of forecast performance due to cloud conditions [21]. Thus, we look into the more informative prediction error for each CMF class. Therein, we compute the MAE of each CMF class by the persistence and hybrid

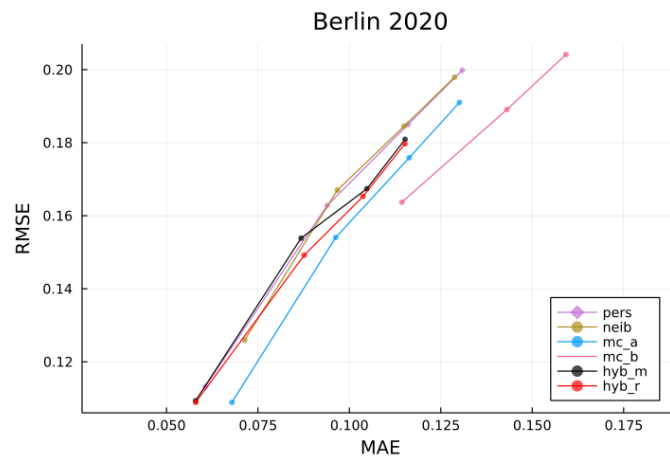


Figure 9. Mean absolute error (MAE) versus root mean square error (RMSE) for Berlin. For one step ahead (+1), the results for both mc_a and mc_b are the same, but only Variant a is visualized to avoid redundancy and/or overlapping. The same applies to the following figures. The results for the 3 other cities can be found in Figure A3 in the Appendix.

approaches for one to four time steps ahead (Figure 10). The MAE of each CMF class grows as the lead time increases, as expected. For 15 to 30 minutes ahead, the persistence method has an advantage in the CMF range between 0.3 and 0.4. This arises due to the fact that the hybrid methods considers the prediction errors of the immediate previous one time step, the CMF class of which could be different from the current time step. For further lead time steps, it is almost invariably the hybrid approaches that generate smaller errors across the whole CMF range. For the CMF classes from 0.5 to 0.8, the improvement by the hybrid methods is the most evident, especially for the lead time of one hour. Except for 1 time step ahead, the errors are large for very small CMF classes. At the CMF classes of around 0.8 to 0.9, there is another peak for large MAE, hinting to the difficulty of accurately predicting the cloud state at these levels. In the case of Berlin, the hybrid approach based on RMSE is more accurate than that based on MAE, except for a few CMF classes at around 0.83.

We have identified the effectiveness of the approaches proposed in a retrospective way. The applicability of these methods mainly depends on the current, namely known information. The change in CMF is the difference of CMF between the current and the immediate previous time step(s), which can be considered as known information. For example, we are currently at 9:30, so the CMF change from 9:00 (t-2) to 9:30 (t) is known. Thus, we evaluate the MAE of predicted CMF as a function of each class of CMF change by different approaches for 1 to 4 time steps ahead (Figure 11). The histograms on the background indicate the amount of data points in each class of CMF change from the preceding time step. The persistence method performs the best when the change in CMF is very small, within the range of ± 0.04 . When the CMF change is larger, the majority of the other approaches outperform the persistence one. Among them, the Variant b of MC prediction has the lowest MAE when the CMF changes largely, followed by the Variant a or the hybrid approach in some classes of CMF change. These tendencies are more distinct when the CMF increases from the past to the present than when the CMF change is in the reverse direction. However, note the majority of the CMF changes resides in a very small range clustered around zero, as can be seen from the distribution of the CMF changes.

After the evaluation of CMF for the city of Berlin, we apply the hybrid method based on RMSE to three other cities introduced in the Section 2. The individual transition matrix is computed for each city. Table 4 lists the Pearson correlation coefficients of predicted CMF by the hybrid method based on RMSE with the actual time series for four lead time steps for four cities in 2020. The correlation is the highest for 1 time step ahead, featuring correlation coefficients larger than 0.9. As the lead time increases, the correlation coefficients decrease to around 0.75. Among the cities considered, the hybrid approach yields the best result for

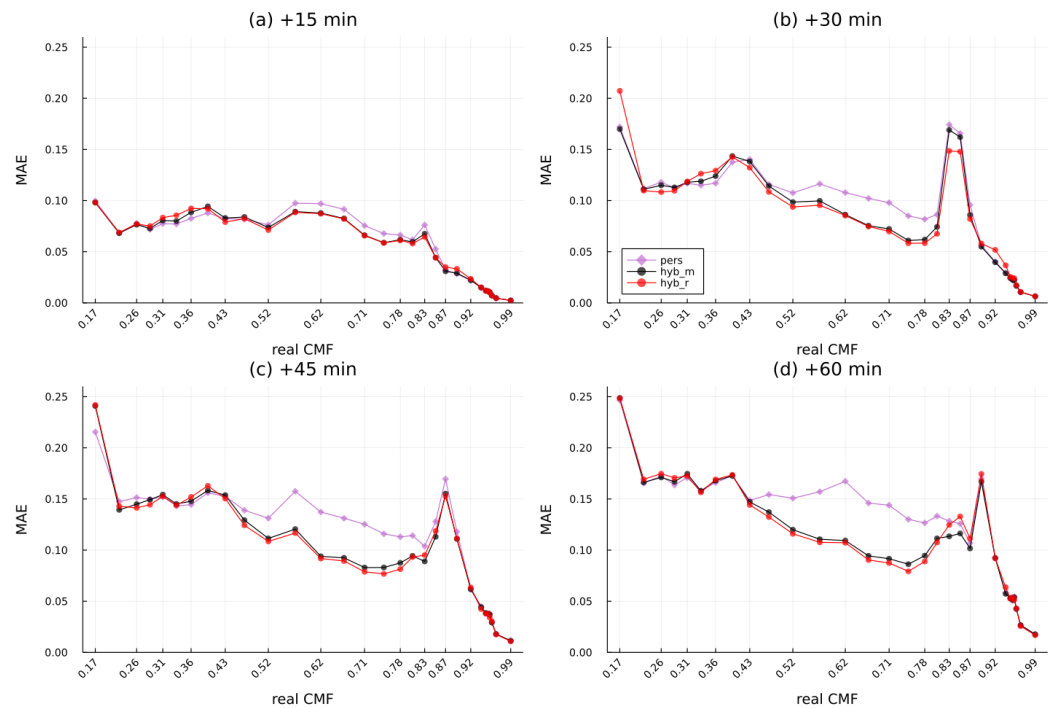


Figure 10. MAE for each CMF class for (a) 15 minutes, (b) 30 minutes, (c) 45 minutes and (d) 60 minutes ahead in 2020 for Berlin. The results for the 3 other cities can be found in Figures A4, A5 and A6 in the Appendix.

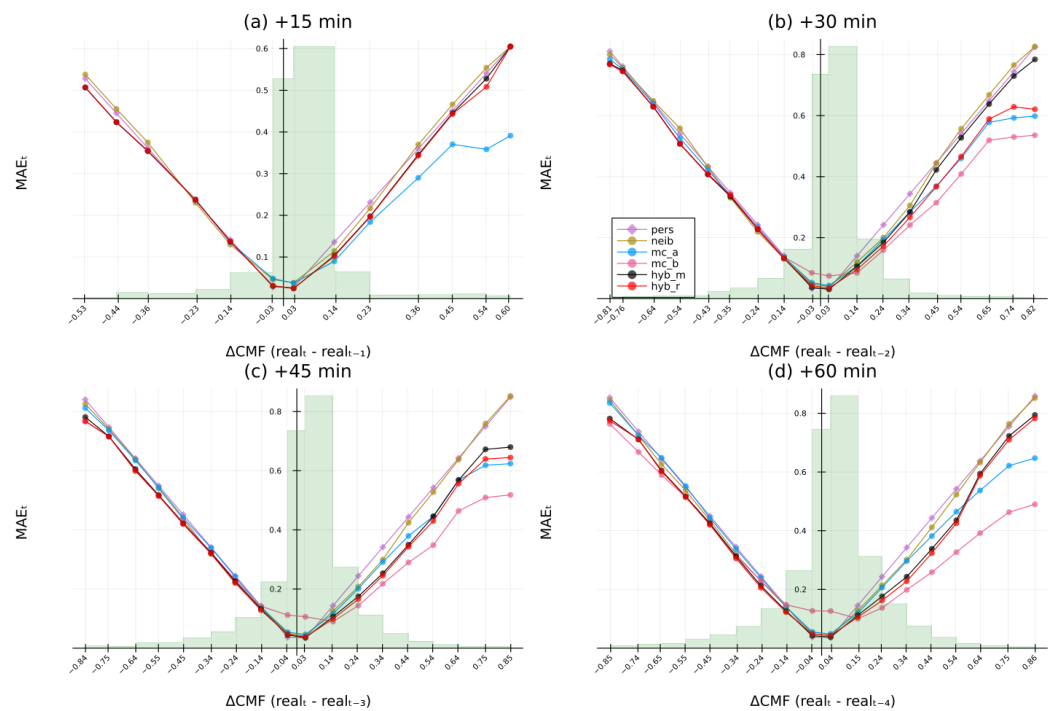


Figure 11. MAE of CMF change for (a) 15 minutes, (b) 30 minutes, (c) 45 minutes and (d) 60 minutes ahead for Berlin. The horizontal ticks are the means of every class of CMF change, where the change classes are equally-spaced with an interval of 0.1, e.g., $[-0.1, 0)$ and $[0, 0.1)$. The results for the 3 other cities can be found in Figures A7, A8 and A9 in the Appendix.

Bucharest for each lead time step, even though Bucharest is not the city with the lowest or the most variable CMF (see Table 1). For Helsinki, the hybrid method works properly from one to four time steps ahead. In the sunnier city Athens, the hybrid method's prediction is comparably good for at least 15 or 30 minutes ahead.

Table 4. Correlation coefficients of predicted CMF by the *hyb_r* with the actual time series for four cities for four lead time steps (+1 to +4) in 2020.

City	+1	+2	+3	+4
Athens	0.9176	0.8406	0.7807	0.7574
Bucharest	0.9547	0.8892	0.8365	0.7962
Berlin	0.9146	0.8352	0.7975	0.7609
Helsinki	0.9418	0.8536	0.81	0.7696

4.2. Results of GHI

Following the method described in Section 3.6, we examine the performance of the hybrid approach based on RMSE in predicting GHI. Figure 12 shows the time series of GHI provided by CAMS and predicted with a lead time of 15 minutes using the hybrid approach based on RMSE for Berlin from 2020-01-01 to 2020-01-11. Nighttime hours without sunshine are not included. We are able to observe both curves in high agreement, especially when the absolute GHI values are high. In moments of relatively low GHI values, the hybrid approach based on RMSE tends to overestimate the GHI, as illustrated by the red overshoots during the course.

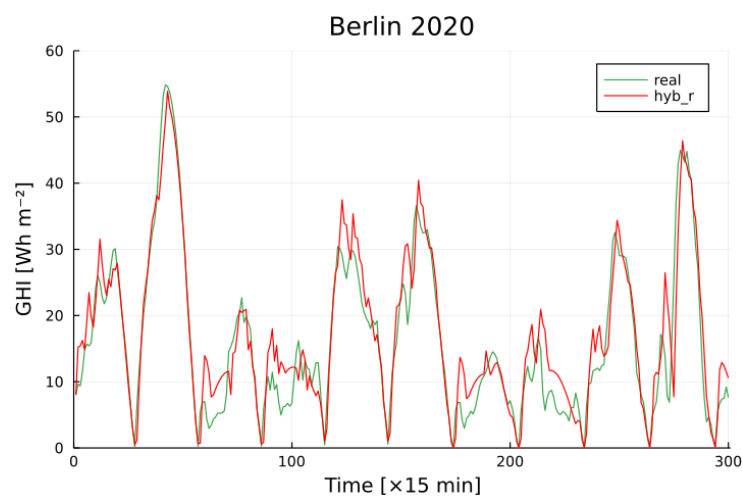


Figure 12. Comparison of actual daily GHI by CAMS (green) and predicted with a lead time of 15 minutes by the hybrid approach based on RMSE (red) for Berlin from 2020-01-01 to 2020-01-11.

When we consider the mean daily insolation (GHI) for each month, the annual course of the prediction for Berlin in 2020 bears reasonably higher resemblance to the one provided by CAMS (Figure 13), since shorter temporal variations of GHI are smoothed out. Especially in months from July to September, the differences between both time series are negligible. In some months with the visible deviation of monthly GHI prediction from the actual time series, there is still an overestimation from the hybrid method, except in April. Positive bias stemming from the MC model could be partially due to the way the MC model is trained in this study: the calculation of GHI or CMF does not take into account the atmospheric mechanisms that could affect cloudiness.

Figure 14 shows the monthly MAE of GHI 1 hour ahead predicted by the persistence and hybrid methods for the four cities in 2020 with their individual vertical axes. In Athens, the performance of the persistence is not necessarily worse than the both hybrid

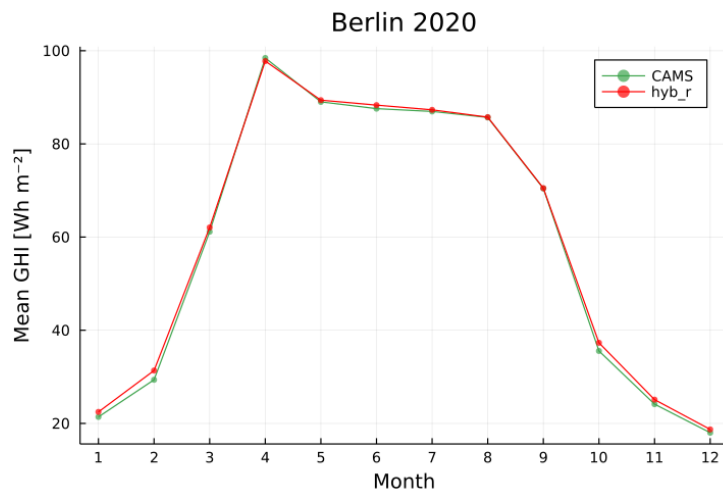


Figure 13. Comparison of actual monthly GHI by CAMS (green) and predicted with a lead time of 15 minutes by the hybrid approach based on RMSE (red) for Berlin in 2020.

methods, only from October to February, when the sunshine duration is shorter in the year. However, the hybrid approach based on MAE has the lowest MAE among them in every month of the year. In Bucharest, both hybrid methods perform similarly, almost always better than the persistence, especially in summer from May to July. The advantage of the hybrid approach over the persistence can be better demonstrated in the city of Berlin, also distinctively during summer, with a monthly MAE reduction of up to 2.5 Wh m^{-2} . In Helsinki, the improvement of hybrid approaches in prediction accuracy is more evident in the last quarter of the year, from October to December.

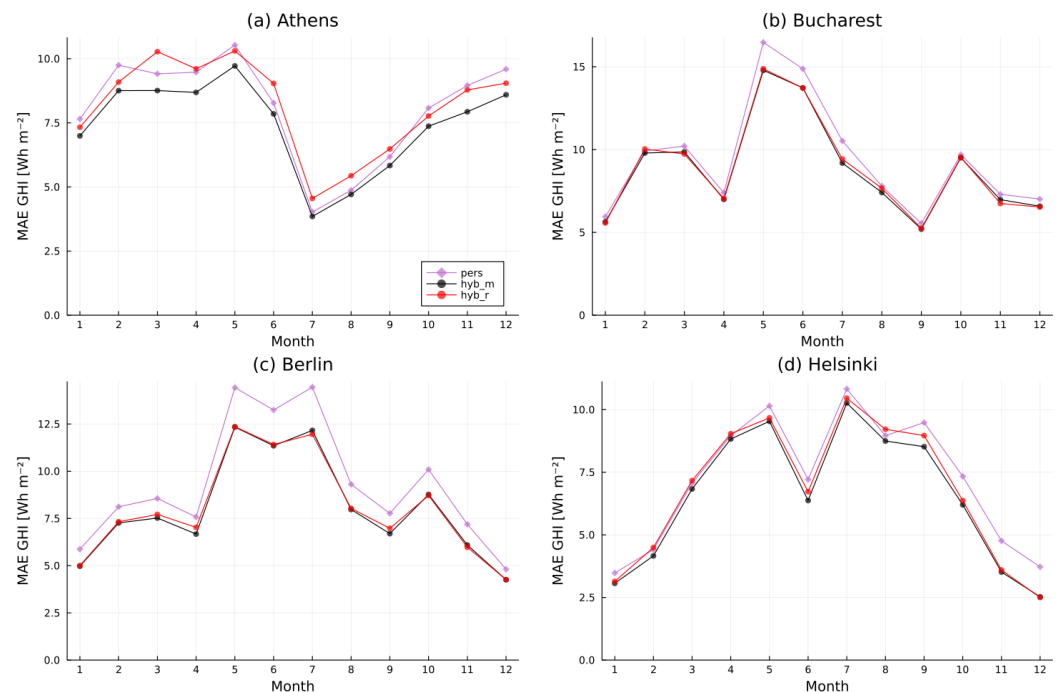


Figure 14. MAE of 1 hour ahead GHI for the persistence and hybrid approaches in 2020 for (a) Athens, (b) Bucharest, (c) Berlin and (d) Helsinki.

To put the results of GHI prediction into perspective, we compare the performance of the hybrid methods with the persistence using the MAE improvement of GHI for each month. This improvement in MAE (in %) is defined as the difference of MAE of GHI by

the hybrid approach ($MAE(\hat{X}_i)$) and MAE of GHI by the persistence approach ($MAE(X_i)$) divided by the actual mean daily GHI in Month i (\bar{X}_i):

$$\Delta MAE_{i,rel} = \frac{MAE(\hat{X}_i) - MAE(X_i)}{\bar{X}_i} \quad (10)$$

Figure 15 shows the monthly relative improvement of the hybrid approach based on MAE compared with the persistence method for four time steps for the four cities. Note that the vertical axis scale has also been adjusted individually to make the results more readable. For one time step ahead in Athens, the hybrid approach is inferior to the persistence approach in predicting GHI. There is some improvement for further time steps, although the magnitude of which is small. The improvement in most of the months is within 2%, and is close to zero in the summer months of July and August. In Bucharest, there is also no improvement for one time step ahead except for the months from May to July with small magnitude. The improvement for the majority of the months are smaller than 2%. In Berlin, on the other hand, we observe an improvement from one to four time steps ahead across all months, except for 15 minutes ahead in February. For a lead time shorter than 45 minutes, the fluctuation of improvement throughout the year is relatively small. However, when predicting one hour ahead, the improvement is apparent during the winter months of October to January, even reaching nearly 5% in November. In Helsinki, there is some general improvement to confirm. In December, we even observe higher than 10% of improvement for the lead time of 45 and 60 minutes. However, this is mainly due to the original low monthly GHI in Helsinki.

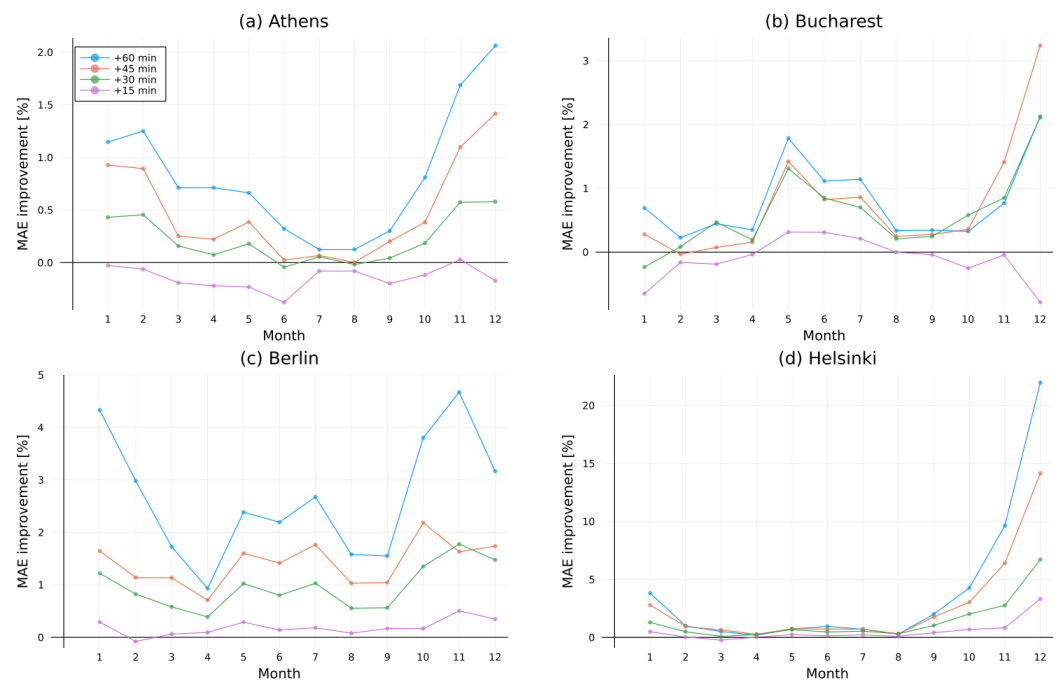


Figure 15. Relative improvement of hybrid approach in 2020 for (a) Athens, (b) Bucharest, (c) Berlin and (d) Helsinki.

5. Conclusions

The accuracy of the solar radiation prediction is subject to the location and forecast horizon considered. For locations with relatively high variability in cloudiness, such as Berlin in this study, the persistence approach does not work satisfactorily for a longer period of time. This is also linked with our motivation to bring up better methodology for the solar nowcasting. In this study, we present a hybrid model for solar radiation nowcasting, which combines the persistence ensemble, the Markov chain model and neighbor inference

through cloud movement. In the GHI prediction for Berlin, the reduction of monthly mean error amounts to 2.5 Wh m^{-2} and the relative improvement reaches nearly 5% compared with the traditional persistence method. In the other three cities, accuracy improvement of various degrees can also be observed.

We acknowledge that the data involved in this study is based on outputs from the models rather than true measurements, which has its own limitations of unable to reflect the reality to 100%. However, note that the proposed model undoubtedly could also be applied to ground-based measurement data. On the other hand, the use of modelled clear-sky radiation is nevertheless inevitable for the prediction of GHI from CMF without a modeling simulation involvement.

The presented approach has the advantage that its straightforward model configuration does not require the input of other variables, such as wind speed or the CMV from the all-sky imagers, both of which can introduce multiple uncertainties. Furthermore, the computational time required for prediction at one single site is trivial. Our approach also adjusts the transition matrix according to the input data for the MC prediction, thus, it is a generic model that can be readily applied to different geographical locations where the CAMS data are available. Therefore, one prospect of the hybrid model is to upscale the application to e.g. the pan-European domain to build an efficient network for solar radiation nowcasting. The presented method can be the base, or part, of hybrid approaches including additional, satellite based cloud information and/or ground measurements (cloud cameras, solar irradiance instruments, other atmospheric parameters measurements) towards improving solar forecasting.

Author Contributions: Conceptualization and methodology, S.K., X.H. and K.P.; software and formal analysis, X.H. and K.P.; data curation, visualization, validation and writing—original draft preparation, X.H.; writing—review and editing, X.H., K.P. Y.-M.S.-D., S.K.; supervision and funding acquisition, S.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by “European Commission EuroGEO Showcases: Applications Powered by Europe” project (grant no. 820852).

Acknowledgments: Data from CAMS radiation service are accessed through Transvalor SoDA at <http://www.soda-pro.com/web-services/radiation/cams-radiation-service> (last access: 10 November 2021).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

Appendix A Appendix

Table A1. Correlation coefficients of predicted with actual cloud modification factor (CMF) time series for 4 lead time steps (+1 to +4) in 2020 for Athens.

Method	+1	+2	+3	+4
pers	0.9157	0.8323	0.7867	0.7509
mc_a	0.9164	0.8394	0.7929	0.7587
mc_b	0.9164	0.816	0.7498	0.6981
neib	0.8332	0.7626	0.7329	0.7065
hyb_m	0.9174	0.8475	0.8069	0.7821
hyb_r	0.9176	0.8406	0.7807	0.7574

References

1. Lefèvre, M.; Oumbe, A.; Blanc, P.; Espinar, B.; Gschwind, B.; Qu, Z.; Wald, L.; Schroedter-Homscheidt, M.; Hoyer-Klick, C.; Arola, A.; et al. McClear: a new model estimating downwelling solar radiation at ground level in clear-sky conditions. *Atmospheric Measurement Techniques* **2013**, *6*, 2403–2418. doi:10.5194/amt-6-2403-2013.

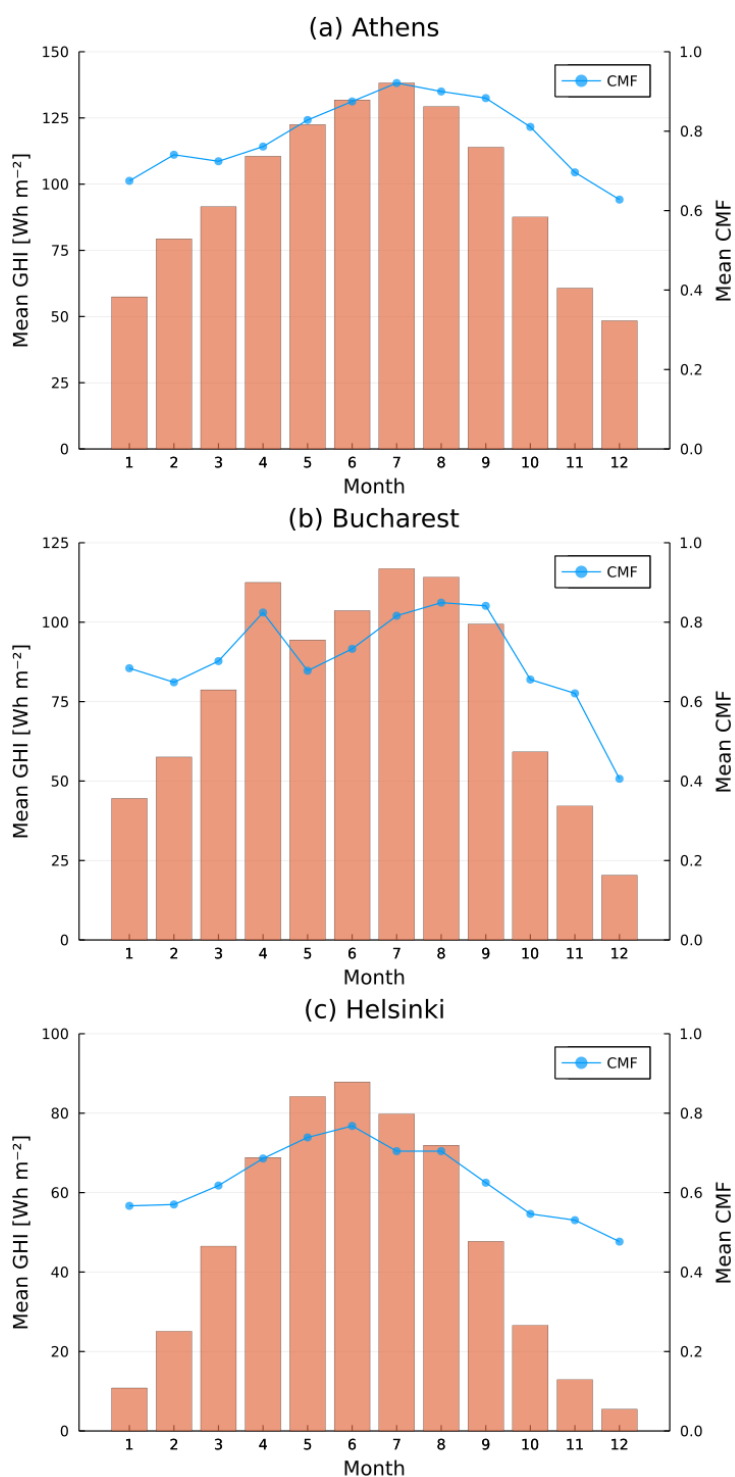


Figure A1. Daily mean global horizontal irradiation (GHI) for each month (orange bars, left axis) and monthly mean cloud modification factor (CMF; blue curves, right axis) in 2020 for (a) Athens, (b) Bucharest and (c) Helsinki.

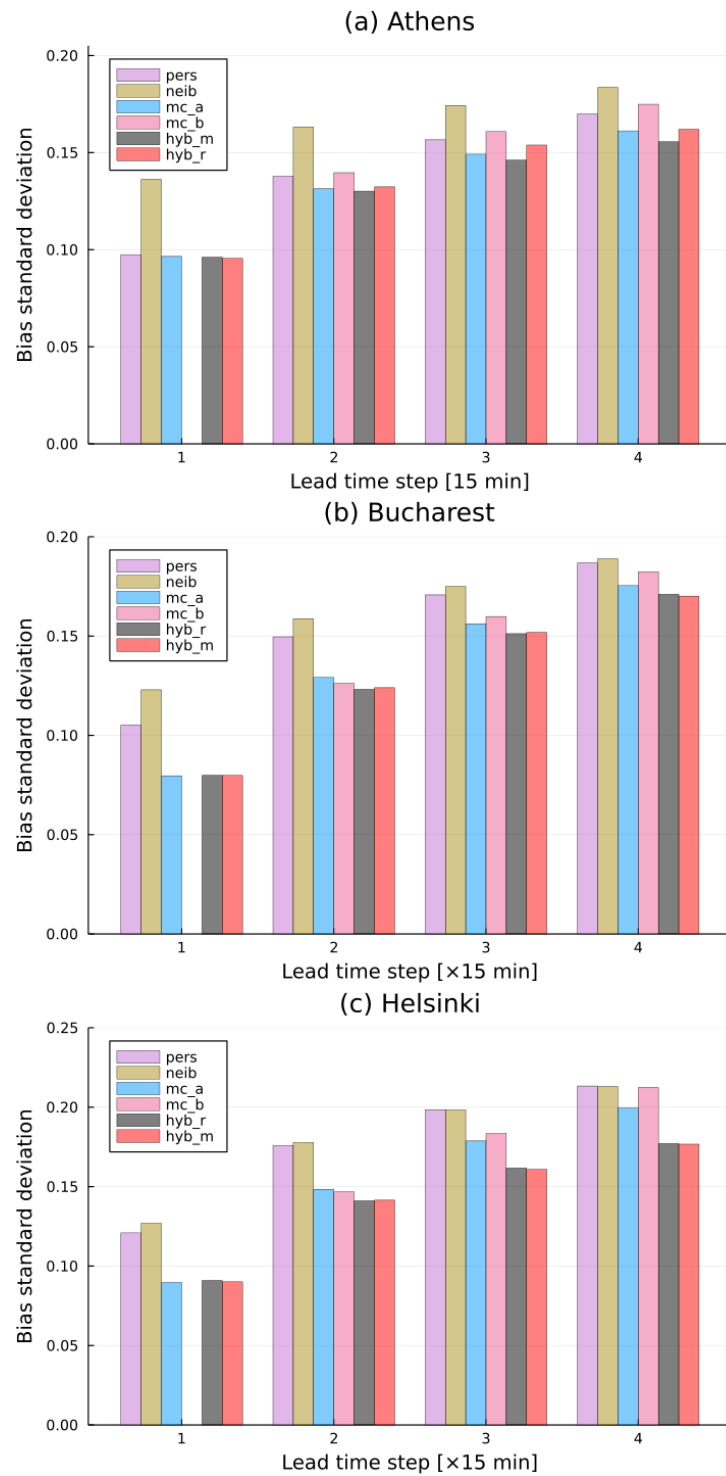


Figure A2. Standard deviation of cloud modification factor (CMF) prediction bias for (a) Athens, (b) Bucharest and (c) Helsinki.

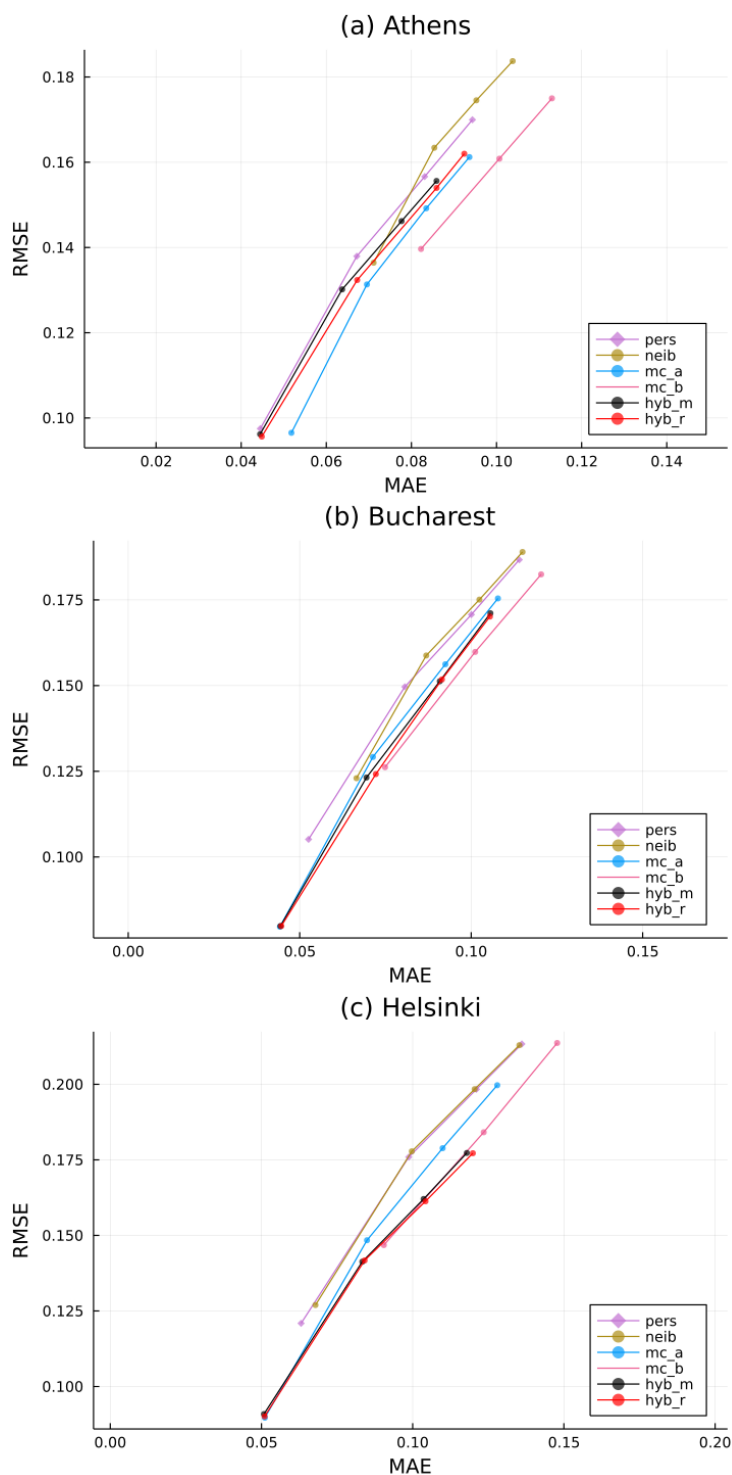


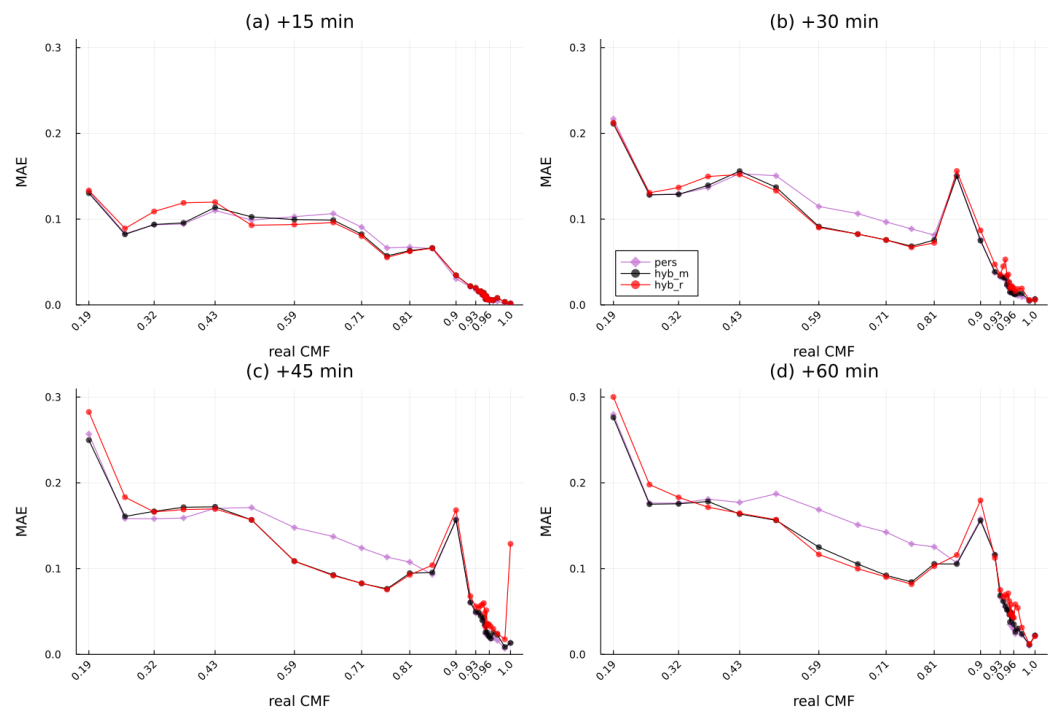
Figure A3. Mean absolute error (MAE) versus root mean square error (RMSE) for (a) Athens, (b) Bucharest and (c) Helsinki.

Table A2. Correlation coefficients of predicted with actual CMF time series for 4 lead time steps (+1 to +4) in 2020 for Bucharest.

Method	+1	+2	+3	+4
pers	0.9226	0.8457	0.8023	0.7663
mc_a	0.955	0.8811	0.8281	0.7854
mc_b	0.955	0.885	0.8136	0.7549
neib	0.8953	0.8281	0.7943	0.7631
hyb_m	0.9547	0.8919	0.8381	0.7954
hyb_r	0.9547	0.8892	0.8365	0.7962

Table A3. Correlation coefficients of predicted with actual CMF time series for 4 lead time steps (+1 to +4) in 2020 for Helsinki.

Method	+1	+2	+3	+4
pers	0.8974	0.7858	0.7319	0.6943
mc_a	0.9424	0.8408	0.7715	0.7189
mc_b	0.9424	0.8413	0.7488	0.6651
neib	0.8878	0.7831	0.7342	0.6982
hyb_m	0.941	0.8551	0.8094	0.7725
hyb_r	0.9418	0.8536	0.81	0.7696

**Figure A4.** MAE for each CMF class for (a) 15 minutes, (b) 30 minutes, (c) 45 minutes and (d) 60 minutes ahead in 2020 for Athens.

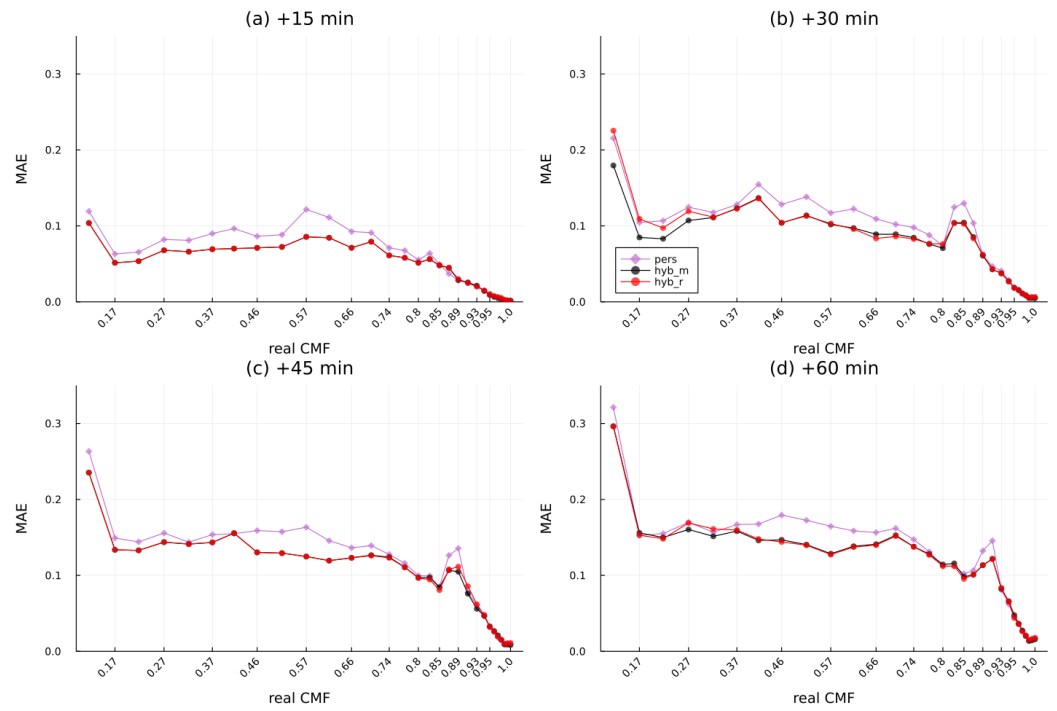


Figure A5. MAE for each CMF class for (a) 15 minutes, (b) 30 minutes, (c) 45 minutes and (d) 60 minutes ahead in 2020 for Bucharest.

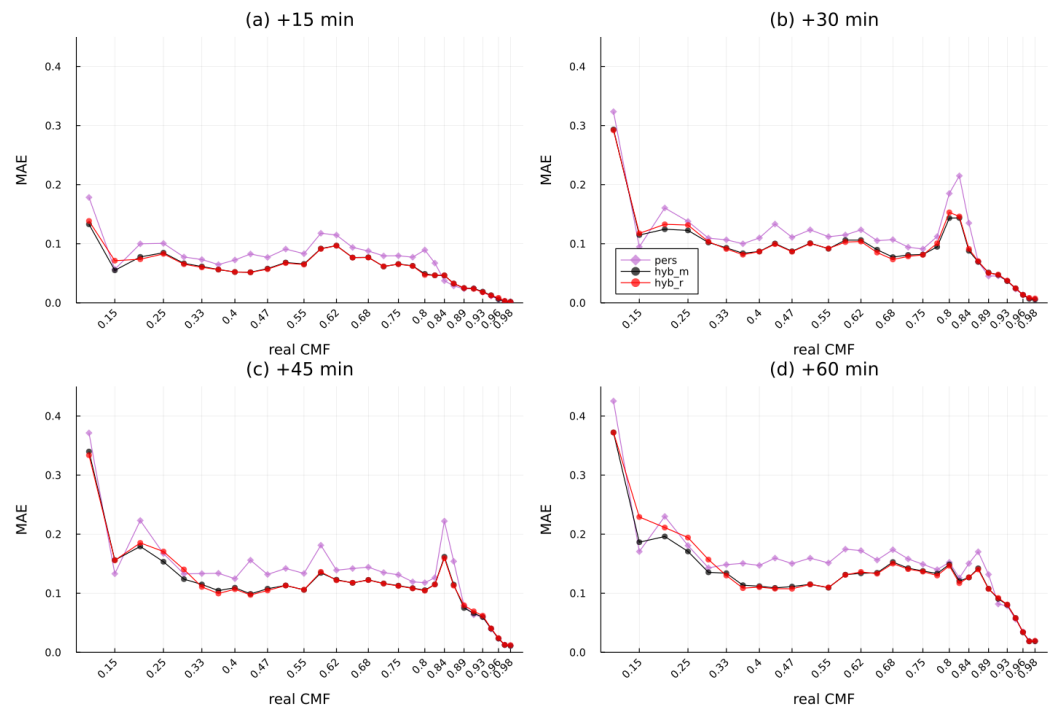


Figure A6. MAE for each CMF class for (a) 15 minutes, (b) 30 minutes, (c) 45 minutes and (d) 60 minutes ahead in 2020 for Helsinki.

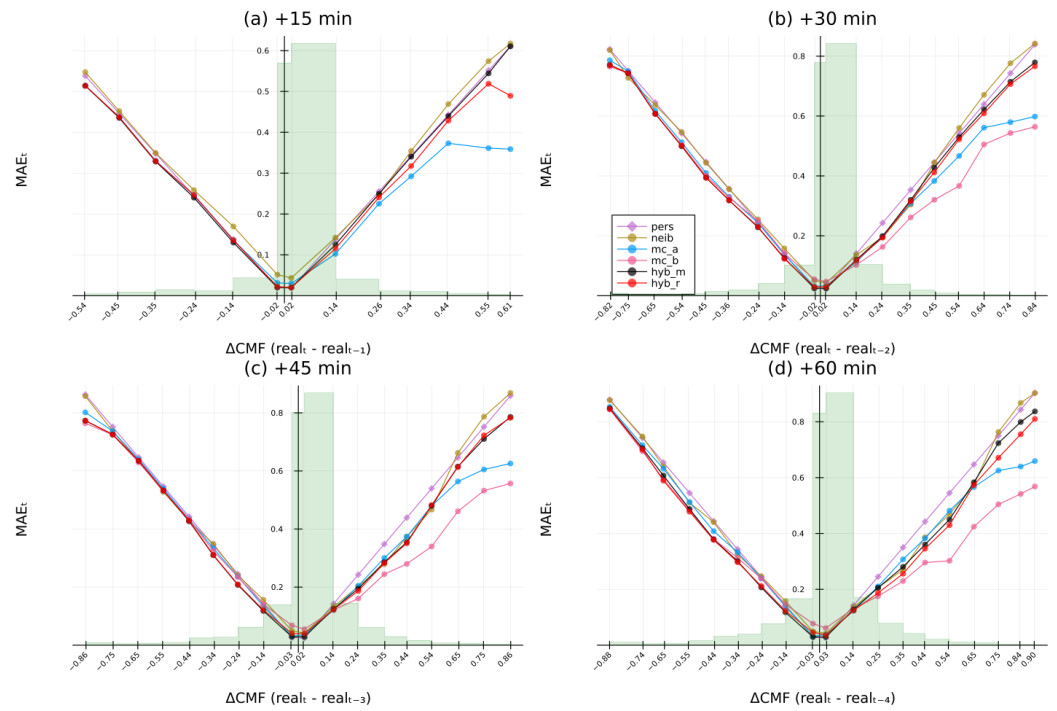


Figure A7. MAE of CMF change for (a) 15 minutes, (b) 30 minutes, (c) 45 minutes and (d) 60 minutes ahead for Athens.

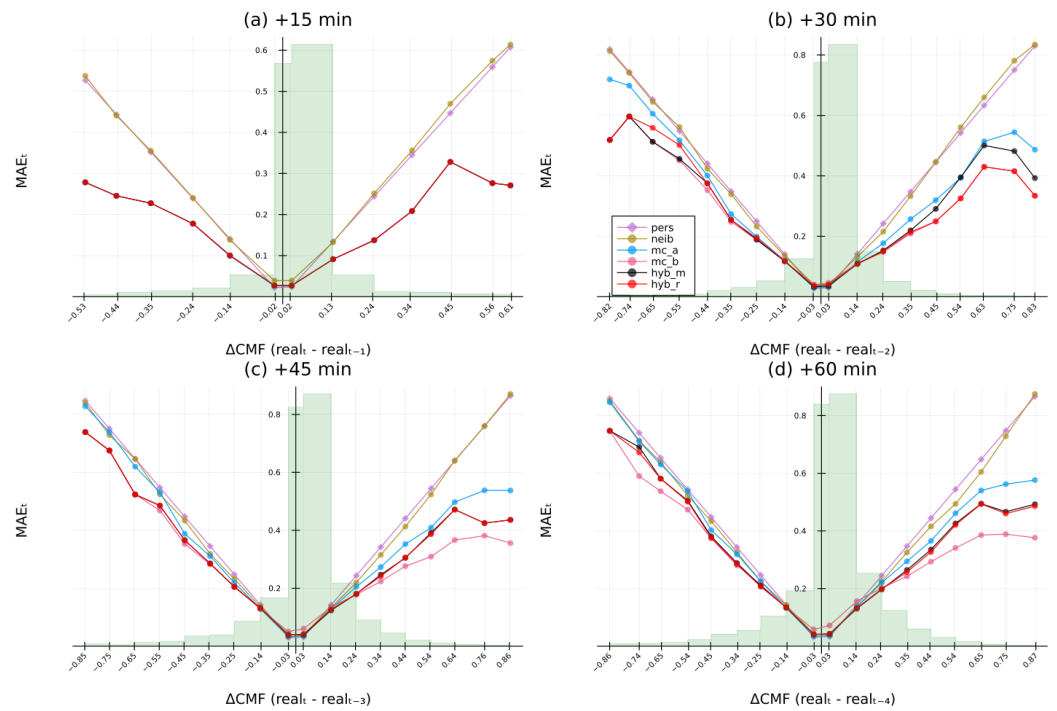


Figure A8. MAE of CMF change for (a) 15 minutes, (b) 30 minutes, (c) 45 minutes and (d) 60 minutes ahead for Bucharest.

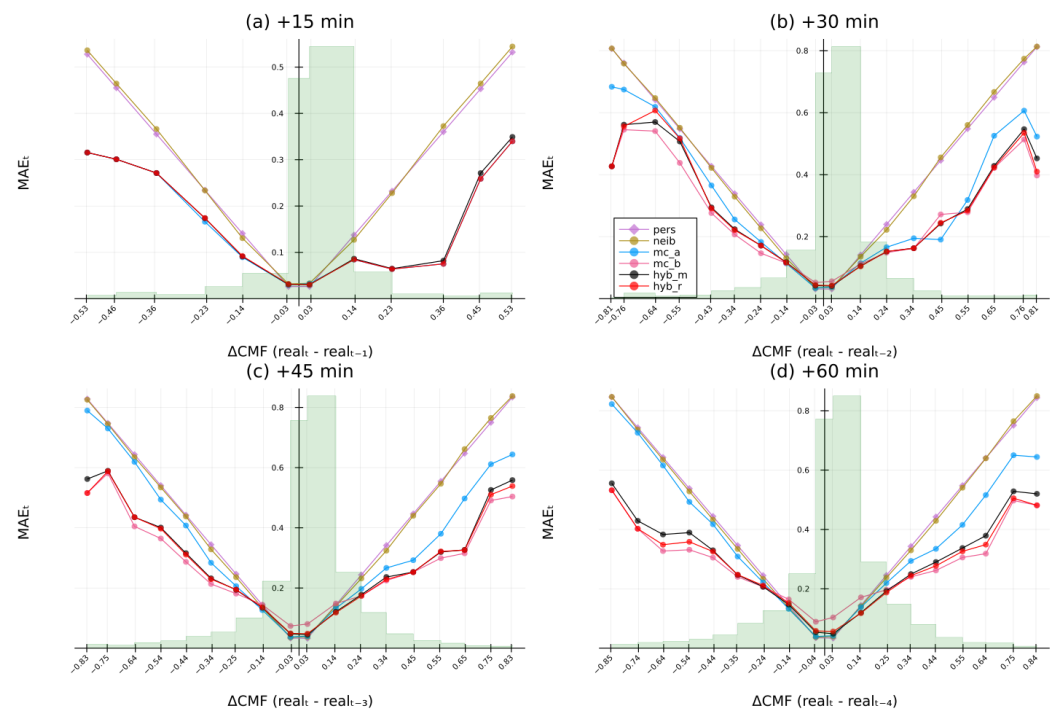


Figure A9. MAE of CMF change for (a) 15 minutes, (b) 30 minutes, (c) 45 minutes and (d) 60 minutes ahead for Helsinki.

2. Mellit, A.; Bengham, M.; Arab, A.H.; Guessoum, A. A simplified model for generating sequences of global solar radiation data for isolated sites: Using artificial neural network and a library of Markov transition matrices approach. *Solar Energy* **2005**, *79*, 469–482. doi:10.1016/j.solener.2004.12.006. 412
3. Ngoko, B.O.; Sugihara, H.; Funaki, T. Synthetic generation of high temporal resolution solar radiation data using Markov models. *Solar Energy* **2014**, *103*, 160–170. doi:10.1016/j.solener.2014.02.026. 413
4. Huang, J.; Korolkiewicz, M.; Agrawal, M.; Boland, J. Forecasting solar radiation on an hourly time scale using a Coupled AutoRegressive and Dynamical System (CARDS) model. *Solar Energy* **2013**, *87*, 136–149. doi:10.1016/j.solener.2012.10.012. 414
5. Diagne, M.; David, M.; Lauret, P.; Boland, J.; Schmutz, N. Review of solar irradiance forecasting methods and a proposition for small-scale insular grids. *Renewable and Sustainable Energy Reviews* **2013**, *27*, 65–76. doi:10.1016/j.rser.2013.06.042. 415
6. Carrière, T.; Amaro e Silva, R.; Zhuang, F.; Saint-Drenan, Y.M.; Blanc, P. A New Approach for Satellite-Based Probabilistic Solar Forecasting with Cloud Motion Vectors. *Energies* **2021**, *14*, 4951. doi:10.3390/en14164951. 416
7. Lorenz, E.; Hurka, J.; Heinemann, D.; Beyer, H.G. Irradiance Forecasting for the Power Prediction of Grid-Connected Photovoltaic Systems. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* **2009**, *2*, 2–10. doi:10.1109/JSTARS.2009.2020300. 417
8. Leutbecher, M.; Palmer, T.N. Ensemble forecasting. *Journal of Computational Physics* **2008**, *227*, 3515–3539. doi:10.1016/j.jcp.2007.02.014. 418
9. Sengupta, M.; Habte, A.; Wilbert, S.; Gueymard, C.; Remund, J. Best Practices Handbook for the Collection and Use of Solar Resource Data for Solar Energy Applications: Third Edition. doi:10.2172/1778700. 419
10. Morf, H. Sunshine and cloud cover prediction based on Markov processes. *Solar Energy* **2014**, *110*, 615–626. doi:10.1016/j.solener.2014.09.044. 420
11. Hocaoglu, F.O.; Serttas, F. A novel hybrid (Mycielski-Markov) model for hourly solar radiation forecasting. *Renewable Energy* **2017**, *108*, 635–643. doi:10.1016/j.renene.2016.08.058. 421
12. Munkhammar, J.; Widén, J. A Markov-chain probability distribution mixture approach to the clear-sky index. *Solar Energy* **2018**, *170*, 174–183. doi:10.1016/j.solener.2018.05.055. 422
13. Munkhammar, J.; van der Meer, D.; Widén, J. Probabilistic forecasting of high-resolution clear-sky index time-series using a Markov-chain mixture distribution model. *Solar Energy* **2019**, *184*, 688–695. doi:10.1016/j.solener.2019.04.014. 423
14. Poggi, P.; Notton, G.; Muselli, M.; Louche, A. Stochastic study of hourly total solar radiation in Corsica using a Markov model. *International Journal of Climatology* **2000**, *20*, 1843–1860. doi:10.1002/1097-0088(20001130)20:14<1843::AID-JOC561>3.0.CO;2-O. 424
15. Bright, J.M.; Smith, C.J.; Taylor, P.G.; Crook, R. Stochastic generation of synthetic minutely irradiance time series derived from mean hourly weather observation data. *Solar Energy* **2015**, *115*, 229–242. doi:10.1016/j.solener.2015.02.032. 425
16. Shepero, M.; Munkhammar, J.; Widén, J. A generative hidden Markov model of the clear-sky index. *Journal of Renewable and Sustainable Energy* **2019**, *11*, 043703. doi:10.1063/1.5110785. 426
17. Urrego-Ortiz, J.; Martínez, J.A.; Arias, P.A.; Jaramillo-Duque, Á. Assessment and Day-Ahead Forecasting of Hourly Solar Radiation in Medellín, Colombia. *Energies* **2019**, *12*, 4402. doi:10.3390/en12244402. 427
18. Soda-service. CAMS Radiation service, 2021. 428

-
19. Qu, Z.; Oumbe, A.; Blanc, P.; Espinar, B.; Gesell, G.; Gschwind, B.; Klüser, L.; Lefèvre, M.; Saboret, L.; Schroedter-Homscheidt, M.; et al. Fast radiative transfer parameterisation for assessing the surface solar irradiance: The Heliosat-4 method. *Meteorologische Zeitschrift* **2017**, *26*, 33–57. doi:10.1127/metz/2016/0781. 445
446
447
20. Gschwind, B.; Wald, L.; Blanc, P.; Lefèvre, M.; Schroedter-Homscheidt, M.; Arola, A. Improving the McClear model estimating the downwelling solar radiation at ground level in cloud-free conditions – McClear-v3. *Meteorologische Zeitschrift* **2019**, *28*, 147–163. doi:10.1127/metz/2019/0946. 448
449
450
21. Vallance, L.; Charbonnier, B.; Paul, N.; Dubost, S.; Blanc, P. Towards a standardized procedure to assess solar forecast accuracy: A new ramp and time alignment metric. *Solar Energy* **2017**, *150*, 408–422. doi:10.1016/j.solener.2017.04.064. 451
452