A measure-correlate-predict approach for optical turbulence (2) using gradient boosting

Pierzyna, Maximilian; Basu, Sukanta; Saathof, Rudolf

DOI
10.1364/PCAOP.2024.PTh1E.3

Publication date
2024

Document Version
Final published version

Citation (APA)
Pierzyna, M., Basu, S., & Saathof, R. (2024). *A measure-correlate-predict approach for optical turbulence (2) using gradient boosting*. Paper presented at Propagation Through and Characterization of Atmospheric and Oceanic Phenomena, pcAOP 2024 - Part of Optica Imaging Congress, Toulouse, France. https://doi.org/10.1364/PCAOP.2024.PTh1E.3

A measure-correlate-predict approach for optical turbulence (C_n^2) using gradient boosting

Maximilian Pierzyna,[1,*] Sukanta Basu,[2] Rudolf Saathof[3]

[1]*Department of Geoscience and Remote Sensing, Delft University of Technology, Delft, The Netherlands*
[2]*Atmospheric Sciences Research Center, University at Albany, Albany, USA*
[3]*Faculty of Aerospace Engineering, Delft University of Technology, Delft, The Netherlands*

[*]*m.pierzyna@tudelft.nl*

Abstract: We present a machine learning-based measure-correlate-predict approach that predicts a multi-year time-series of optical turbulence strength (C_n^2) with high accuracy ($\bar{r} = 0.78$ at 16 locations) based on a single year of in-situ C_n^2 measurements and reanalysis data. © 2024 The Author(s)

Quantifying optical turbulence strength (C_n^2) at a site of interest to astronomy or free-space optical communication (FSOC) typically requires in-situ measurements [1] or mesoscale simulations [2]. Both approaches become expensive or infeasible when C_n^2 needs to be obtained for long periods, such as multiple years. We propose a machine learning-based measure-correlate-predict (ML-MCP) approach as a solution, which expands a single year of measured C_n^2 to a multi-year C_n^2 time series. Our ML-MCP approach utilizes gradient boosting [3,4] (GB) models to correlate the 1-year observation of C_n^2 data to concurrent and collocated ERA5 reanalysis data. The ERA5 reanalysis is a global dataset from the European Centre for Medium-Range Weather Forecasts (ECMWF), providing hourly estimates of a wide range of atmospheric, land, and oceanic variables from 1950 to the present [5]. After the GB models are fitted, arbitrary periods of ERA5 can be used as model inputs to obtain corresponding predicted C_n^2 values. Note that such predictions are only valid for the location where the underlying observations were collected.

For a comprehensive performance assessment of the ML-MCP approach, we utilize the 16 flux stations of the New York State (NYS) Mesonet. These stations are placed across the state of New York and form a diverse network with sites in flat, mountainous, maritime, and urban regions. At each station, C_n^2 is obtained from high-frequency temperature measurements using sonic anemometers at 9 m above ground. Five years (2018 – 2022) of such single-level C_n^2 observations with collocated and concurrent ERA5 data are available for each location. That data is used to train five GB models per site in a round-robin fashion: each model is trained on one year of data and evaluated on the remaining four years. This process is repeated five times until all years are used for training once. We quantify the agreement between the observed and corresponding predicted $\log_{10} C_n^2$ time series of each model employing the Pearson's correlation coefficient r and the root-mean-squared error (RMSE) ε.

These resulting r and ε scores are presented in Fig. 1 in panels (a) and (b), respectively. The scores are similar across years at a single location, i.e. column-wise, demonstrating that there is no strong dependency on the year selected for training. That is encouraging for practical application as the similarity suggests that any available observed year – potentially also archive data – is suitable for temporal extrapolation. The models trained for different sites yield mostly similar performance with little spread around the mean values $\bar{r} = 0.78$ and $\bar{\varepsilon} = 0.42$. Due to the climatological diversity of the network, this low spread demonstrates that our ML-MCP method is applicable in various topographic and meteorological settings. Two groups of climatologically similar sites (cf. grey brackets) stand out due to their model scores forming the tails of the performance distributions: the coastal sites (left-most group) perform above average, while the Champlain Valley sites (right-most group) perform lower than average. The Brooklyn station (BLKN, coastal) and the Whitehall station (WHIT, Champlain Valley) are assessed as representatives of their climatologies in more detail in Fig 2. Comparing the correlation plots (left) in panels (a) and (b) reveals that the lower correlation values for WHIT are due to low C_n^2 values being overestimated. BKLN does not show this tendency. The overestimation is also visible in the six randomly selected 7-day windows (right) where observed (black) and predicted (red) $\log_{10} C_n^2$ are compared. Again, the WHIT predictions tend to miss the low C_n^2 conditions compared to the BKLN time series, thus resulting in larger errors and lower correlation. Since WHIT's climatological sister station, CHAZ, shows a similar behavior, we assume the trained models systematically miss some local processes. That is unsurprising because ERA5 has a horizontal resolution of only ~ 30 km, which could be too coarse to capture the relevant weather phenomena in a complex valley environment such as the Champlain Valley. On the other hand, the coastal climate seems to be better represented in ERA5, leading to consistently better C_n^2 model performance at the three coastal sites.

Overall, our C_n^2 ML-MCP approach performs well and is shown to work reliably for various locations and with different years of training data. Our work enables the accurate temporal extrapolation of an observed 1-year C_n^2

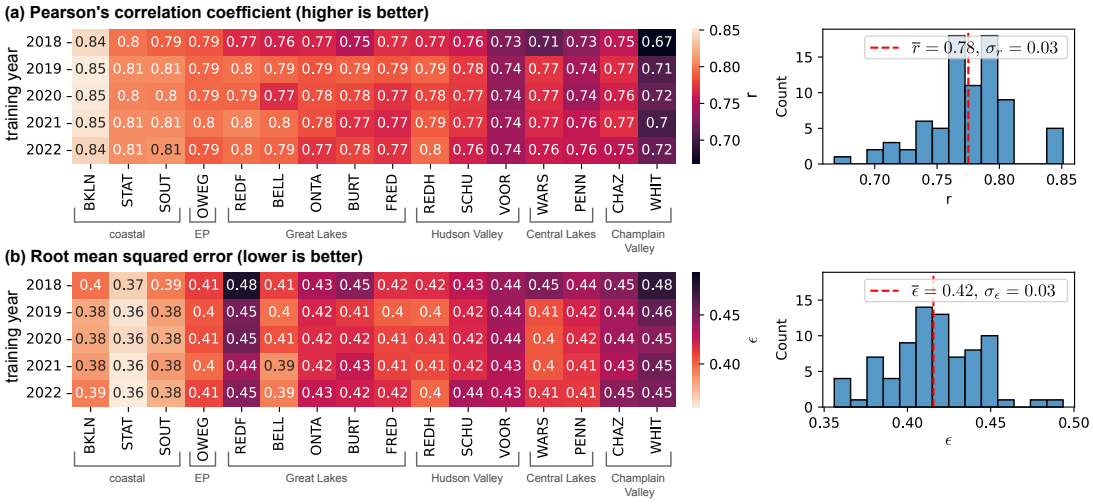

Fig. 1. Overview of ML-MCP performance for all sites. Columns represent stations grouped by climatological similarity (grey brackets), and rows correspond to the year selected for training. The scores in each cell – Pearson correlation coefficient r and root mean squared error ε – reflect the model's performance on all other years that are not used for training. The histograms present the overall distribution of the metrics.

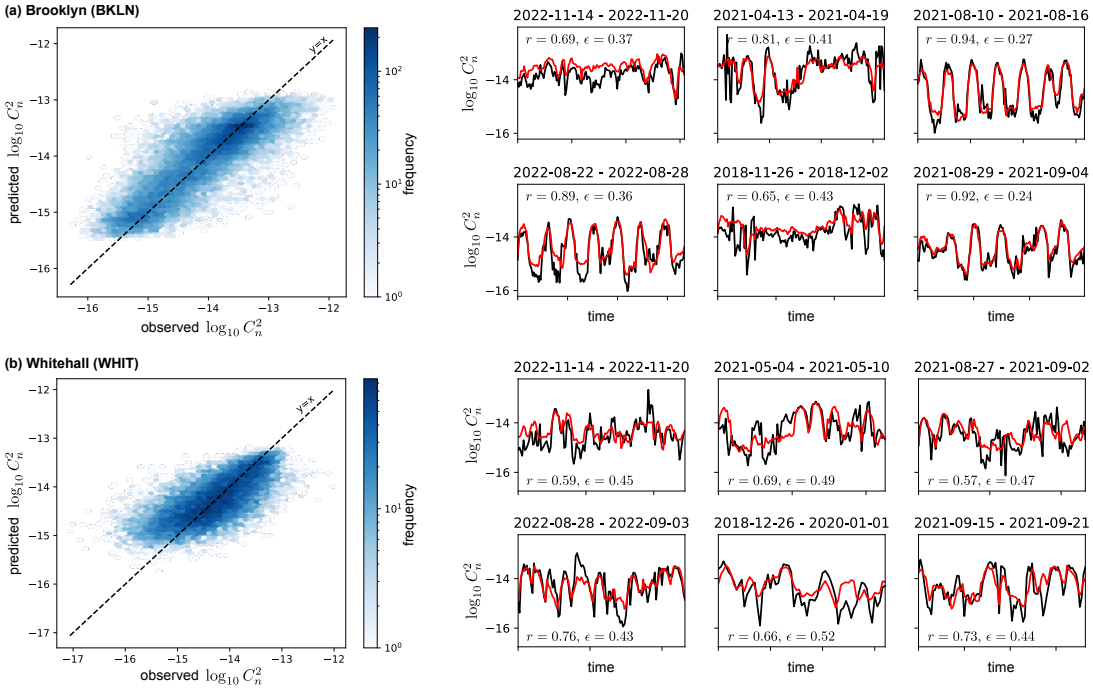

Fig. 2. Correlation plot (left) and time series plots (right) for sites with (a) highest and (b) lowest performance. Time series plots show 16 randomly drawn samples of 7 consecutive days comparing observed (black) and predicted (red) $\log_{10} C_T^2$.

time series to multiple years, which we believe is beneficial for a range of astronomy or FSOC applications.

Acknowledgement

We are grateful to the New York State (NYS) Mesonet for providing the data for this study. MP is funded by the project FREE - Optical Wireless Superhighways: Free photons (at home and in space) (with project number P19-13) of the research programme TTW-Perspectief which is (partly) financed by the Dutch Research Council (NWO).

References

1. J. C. Wyngaard, Y. Izumi, and S. A. Collins, "Behavior of the Refractive-Index-Structure Parameter near the Ground*," J. Opt. Soc. Am. **61**, 1646 (1971).
2. E. Masciadri, J. Vernin, and P. Bougeault, "3D mapping of optical turbulence using an atmospheric numerical model - I. A useful tool for the ground-based astronomy," Astron. Astrophys. Suppl. Ser. **137**, 185–202 (1999).
3. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining,* (Association for Computing Machinery, New York, NY, USA, 2016), KDD '16, pp. 785–794.
4. G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *Advances in Neural Information Processing Systems,* vol. 30 (Curran Associates, Inc., 2017).
5. H. Hersbach, B. Bell, P. Berrisford, S. Hirahara, A. Horányi, J. Muñoz-Sabater, J. Nicolas, C. Peubey, R. Radu, D. Schepers, A. Simmons, C. Soci, S. Abdalla, X. Abellan, G. Balsamo, P. Bechtold, G. Biavati, J. Bidlot, M. Bonavita, G. Chiara, P. Dahlgren, D. Dee, M. Diamantakis, R. Dragani, J. Flemming, R. Forbes, M. Fuentes, A. Geer, L. Haimberger, S. Healy, R. J. Hogan, E. Hólm, M. Janisková, S. Keeley, P. Laloyaux, P. Lopez, C. Lupu, G. Radnoti, P. Rosnay, I. Rozum, F. Vamborg, S. Villaume, and J.-N. Thépaut, "The ERA5 global reanalysis," Q. J. Royal Meteorol. Soc. **146**, 1999–2049 (2020).