Π-ML: a dimensional analysis-based machine learning parameterization of optical turbulence in the atmospheric surface layer

Pierzyna, M.; Saathof, R.; Basu, S.

DOI
10.1364/OL.492652

Publication date
2023

Document Version
Final published version

Published in
Optics Letters

Optics Letters

Π-ML: a dimensional analysis-based machine learning parameterization of optical turbulence in the atmospheric surface layer

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

[1]Faculty of Civil Engineering and Geosciences, Delft University of Technology, Delft, Netherlands
[2]Faculty of Aerospace Engineering, Delft University of Technology, Delft, Netherlands
[3]Atmospheric Sciences Research Center, University at Albany, Albany, New York 12226, USA
*m.pierzyna@tudelft.nl

Received 21 April 2023; revised 25 July 2023; accepted 29 July 2023; posted 31 July 2023; published 18 August 2023

Turbulent fluctuations of the atmospheric refraction index, so-called optical turbulence, can significantly distort propagating laser beams. Therefore, modeling the strength of these fluctuations (C_n^2) is highly relevant for the successful development and deployment of future free-space optical communication links. In this Letter, we propose a physics-informed machine learning (ML) methodology, Π-ML, based on dimensional analysis and gradient boosting to estimate C_n^2. Through a systematic feature importance analysis, we identify the normalized variance of potential temperature as the dominating feature for predicting C_n^2. For statistical robustness, we train an ensemble of models which yields high performance on the out-of-sample data of $R^2 = 0.958 \pm 0.001$.

https://doi.org/10.1364/OL.492652

Free-space optical communication (FSOC) between satellites and ground or between multiple ground terminals is among the emerging applications in which an optical beam propagates through the atmosphere. FSOC can have a major societal impact, increasing data throughput, data security, and global internet coverage while potentially reducing the cost per bit per second [1]. However, some challenges need to be addressed; in addition to precipitation, clouds, fog, and aerosol scattering, turbulent fluctuations of the atmospheric refractive index form a major source of disturbance [2]. The strength of these fluctuations—called optical turbulence—is quantified by the refractive index structure parameter C_n^2. Good knowledge about the behavior of C_n^2 in diverse locations and meteorological conditions is required to design and deploy reliable future FSOC links. However, measuring C_n^2 is difficult and typically needs elaborate post-processing of high-frequency observations [3]. As a result, a wide range of empirical C_n^2 models and parameterizations have emerged, which aim to relate C_n^2 to more easily obtainable variables [4]. Conventional physics-based C_n^2 parametrizations typically make use of Monin–Obukhov similarity theory (MOST) [5] and associated empirically determined similarity relationships. One of the earliest parameterizations was proposed by [3] and utilizes turbulent fluxes to estimate C_n^2. Several other competing formulations exist (refer to [6] for a comprehensive review). Recently, multiple studies [7–10] showed that machine learning (ML) models can be used to parameterize C_n^2 based on routinely available meteorological inputs. These ML approaches parameterize the underlying physical processes from data through sophisticated regression, but they do not explicitly incorporate physical knowledge. In this Letter, we propose an alternative physics-inspired ML framework. We present Π-ML, a dimensional analysis-based ML framework, which strives to improve conventional MOST-based surface layer parameterizations with the power of ML. We use dimensional analysis (DA) constrained with domain knowledge to expand the set of traditional MOST variables and an ensemble of gradient-boosting ML regression models to learn similarity relationships from observations. In DA, the relevant dimensional variables of a physical process are combined into non-dimensional groups which describe that process equally well [11]. DA is compelling to use in practice because the non-dimensional variables enable us to combine observational data from different field campaigns around the world. More importantly, when using ML, DA can change the extrapolation problem in dimensional variables to an interpolation problem in non-dimensional variables [12].

To investigate the strengths and weaknesses of the proposed methodology, we use measurements collected during a seeing study at the Mauna Loa Observatory (MLO) on the island of Hawai'i. The MLO study was conducted by the National Center for Atmospheric Research (NCAR) from 9 June 2006 to 8 August 2006 (~8 weeks). The dataset contains measurements of mean meteorological quantities, turbulent fluxes, and turbulent variances obtained from three sonic anemometers deployed at ca. 6-m, 15-m, and 25-m altitude. The C_n^2 values were estimated by NCAR via inertial-range scaling of temperature spectra [13]. We compute two gradients from the mean horizontal wind components $\overline{u}$ and $\overline{v}$ and the mean potential temperature $\overline{\theta}$: mean wind shear $S = \sqrt{(\partial\overline{u}/\partial z)^2 + (\partial\overline{v}/\partial z)^2}$ and mean potential temperature gradient $\Gamma = \partial\overline{\theta}/\partial z$. Atmospheric turbulence is modulated by thermal buoyancy and wind shear effects, which are captured by the sensible heat flux $\overline{w'\theta'}$ and the friction velocity $u_* = (\overline{u'w'}^2 + \overline{v'w'}^2)^{1/4}$, respectively, where $\overline{u'w'}$ and $\overline{v'w'}$ are momentum flux components. Additionally, we incorporate the

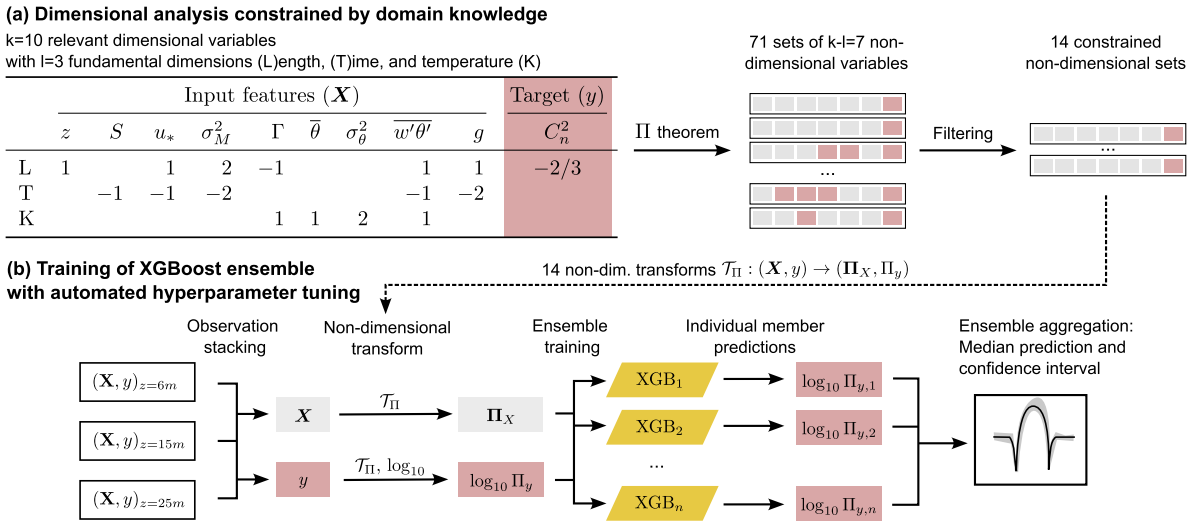

Fig. 1. Our Π-ML methodology consists of two components. (a) The dimensional analysis based on the Buckingham Π theorem combines observed dimensional variables into Π sets of normalized non-dimensional variables. (b) These sets are used to transform the observed data into a stacked non-dimensional dataset to train an ensemble of XGBoost regression models.

variances of potential temperature and horizontal wind magnitude, σ_θ^2 and $\sigma_M^2 = \sigma_u^2 + \sigma_v^2$. The altitude z serves as length scale because we aim for a surface layer C_n^2 parameterization, but to model OT at higher altitudes, the suitable length scale is expected to differ. All relevant variables forming the input for our Π-ML methodology are summarized in the table of Fig. 1(a) with their respective fundamental dimensions. Earth's gravitational acceleration $g = 9.81\,\mathrm{m\,s^{-2}}$ is also included because it is required for the atmospheric force balance. Given the dry atmospheric conditions at the MLO sites, moisture variables were ignored. To later assess the C_n^2 estimation performance of the trained model, the first two weeks of July 2006 are set aside as test data. Although using data out of the middle as test data might seem unconventional, it is used so that the ML models can capture the seasonal change from June to August (see Section 2 of the Supplement 1 for details).

The two key components of our proposed Π-ML methodology are illustrated in Fig. 1: the DA constrained with domain knowledge in panel (a) and the ensemble of gradient-boosting ML models, which perform regression on the stacked, non-dimensionalized observations in panel (b). We set off with the table in Fig. 1(a) and the Buckingham Π theorem [11], popular in DA. The theorem states that our $k = 10$ dimensional variables with their $l = 3$ fundamental dimensions (length, time, temperature) can be expressed as a set of $(k - l) = 7$ independent non-dimensional Π groups. Multiple options exist to form these sets, so we employ the Π theorem implementation of [14], which generates 71 sets with 7 Π groups each. Using domain knowledge, we conceive three constraints to reduce the number of sets from 71 to 14. First, each set can only contain one dependent Π group that is a function of C_n^2 [cf. pink highlights in Fig. 1(a)]. All other Π groups should only be functions of the independent dimensional variables $\boldsymbol{X}$. Second, C_n^2 and its normalized variant Π_y vary over multiple orders of magnitude, so the ML models are trained on $\log_{10}\Pi_y$. Since the logarithm is not defined for negative arguments, only Π sets where Π_y is strictly positive are valid. Third, the dimensional variables Γ and $\overline{w'\theta'}$ can be positive and negative, so raising them to fractional or even-integer powers can result in complex values or a loss of sign. Therefore, valid Π sets cannot contain such expressions. Each of the 14 constrained Π sets is used to scale and non-dimensionalize the dimensional observations $\boldsymbol{X}$ and $y = C_n^2$ to yield $\boldsymbol{\Pi}_X$ and Π_y, respectively, as illustrated in Fig. 1(b). The non-dimensionalized observations from all three levels can be stacked into a combined dataset from which ML learns the non-dimensional black box similarity relationship $f(\boldsymbol{\Pi}_X) \approx \log_{10}\Pi_y$. For each Π set, we train one ensemble of $n = 25$ member models to make robust C_n^2 predictions with uncertainty estimates using the gradient boosting algorithm XGBoost (XGB) and the AutoML library FLAML [15]. FLAML performs time-constrained hyperparameter tuning of the XGB models using 5-fold cross-validation. For each ensemble member, FLAML was given a 10-minute time budget on 8 cores of a 3-GHz Intel Xeon E5-6248R CPU. Such a time-constrained optimization is crucial to keep the overall training costs reasonable (~34 core hours per ensemble). We employ the Monte Carlo resampling strategy to generate a different 4-week subset of the 6-week training data for each member. Two non-overlapping sets of seven consecutive days are randomly removed from the training data, so each subset covers different meteorological conditions. As depicted in Fig. 1(b), each of the n trained members produces a prediction that is robustly aggregated into an ensemble prediction using the median.

The prediction accuracy and model complexity of each trained Π-ML ensemble is assessed to decide which Π set is best suited for our ML-based parameterization. The root mean squared error (RMSE) $\epsilon = \sqrt{\langle (y - \hat{y})^2 \rangle}$ in the log-space is used to quantify accuracy as the deviation between the observed $\log_{10} C_n^2 = y$ from the test set (July 1–14) and the corresponding Π-ML prediction $\hat{y} = \log_{10} \hat{C}_n^2$. We also evaluate the complexity of the Π sets and their trained ML ensembles. That is essential because ML models should only be as complex as necessary to increase their ability to perform well on new unseen data [16]. One Π set is considered simpler than another set if its Π groups are constructed from fewer dimensional variables. Similarly, one trained ensemble is considered simpler than another one if fewer Π groups are important for the ML prediction, i.e., the modeled

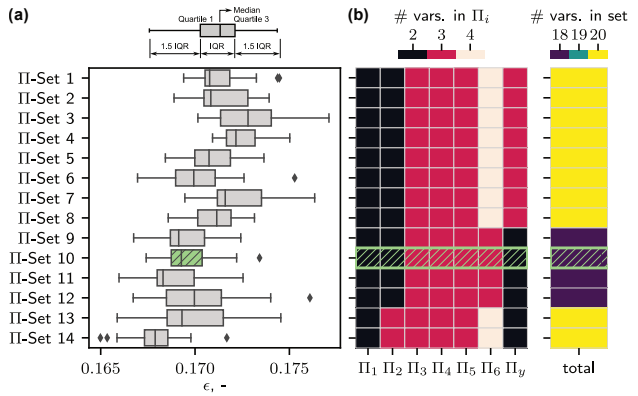

Fig. 2. Comparison of (a) ensemble performance and (b) Π set complexity for our 14 different Π sets, where winning set 10 (green/hatched) balances performance and complexity well.

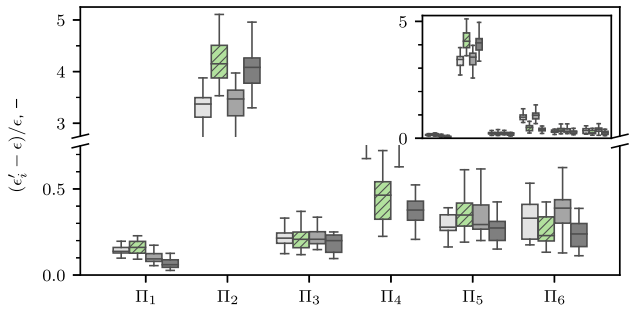

Fig. 3. Importance of non-dimensional Π groups (i.e., Π_1 to Π_6) based on the permutation feature importance strategy. For easy intercomparison, for each non-dimensional Π-group, four boxplots representing ML ensembles corresponding to Π-sets 9, 10, 11, and 12 (left to right) are plotted side-by-side. The best performing ensemble 10 is marked in green (hatched).

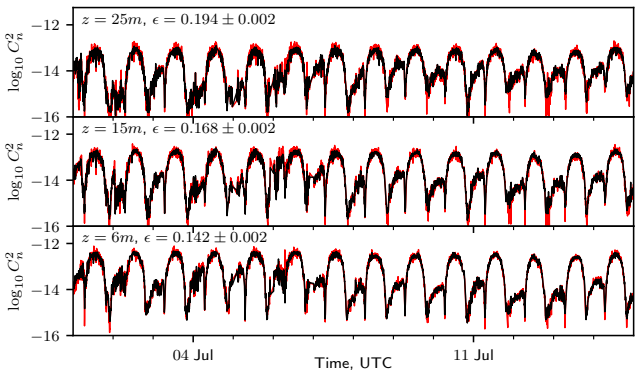

Fig. 4. Median predictions of $\log_{10} C_n^2$ based on test data (black) using the selected Π set 10 ensemble. The observed values (red) are shown for reference.

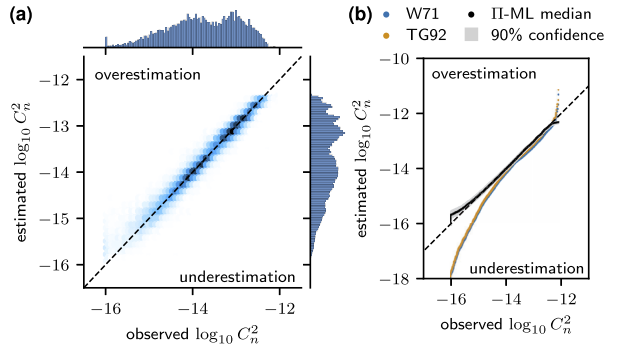

Fig. 5. Correlation histogram and quantile–quantile plot for Π set 10 ensemble showing (a) high correlation ($R^2 = 0.958 \pm 0.001$) and (b) well-captured C_n^2 distributions compared with traditional models from the literature (blue, orange).

C_n^2 is sensitive to fewer input features. The importance of input features of the trained Π-ML models is quantified by the permutation feature importance technique (PFI) [17]. For each feature Π_i, PFI yields a ratio $(\epsilon_i' - \epsilon)/\epsilon$, which describes how the RMSE ϵ_i of a trained model magnifies when the model gets shuffled data for Π_i compared with the baseline RMSE ϵ where the correlation is intact. That means a highly important feature results in a large error magnification.

The performance and complexity of the 14 Π-ML ensembles are shown in Fig. 2. The boxplots in panel (a) display the ϵ distributions for each ensemble. While all ensembles show median RMSEs of the same order of magnitude, Π sets 9 to 14 outperform the others. Panel (b) visualizes complexity through the number of dimensional variables constituting each Π group (left) together with the sum per set (right). This plot reveals that sets 9 to 12 of the well-performing ensembles are the only ones consisting of Π groups formed from no more than three dimensional variables. These four low-error, low-complexity candidates are further assessed based on their PFI score distributions displayed in Fig. 3. Remember that the DA yields different functional expressions for Π_i for each set, which is why each set shows different PFI distributions. The boxplots reveal that Π sets 9 and 11 yield more complex Π-ML ensembles compared with 10 and 12 because they significantly rely on two Π groups (Π_2 and Π_4, see inset) for C_n^2 estimation instead of one (Π_2). Consequently, only sets 10 and 12 remain candidates for our ML-based similarity theory of optical turbulence. From these, we ultimately select Π set 10 because of the lower ϵ spread in Fig. 2(a) with $\Pi_1 = \sigma_M^2/u_*^2$, $\Pi_2 = \overline{\theta}/\sqrt{\sigma_\theta^2}$, $\Pi_3 = (S\,z)/u_*$, $\Pi_4 = \overline{w'\theta'}/(u_*\sigma_\theta)$, $\Pi_5 = (g\,z)/u_*^2$, $\Pi_6 = (\Gamma\,z)/\sigma_\theta$, and $\Pi_y = (C_n^2)^{3/2}\,z$. The expressions for the other 13 Π sets are listed in Supplement 1. The observation that Π_2—the inverse normalized potential temperature variance—is the only dominating feature of our parameterization could have practical implications. First, temperature variances can be measured with thermocouples [18], which are cheaper than sonic anemometers. Second, the low relevance of the gradients (Π_3 and Π_6) indicates that even single-level measurements might be sufficient to estimate C_n^2 accurately. Therefore, our approach might lead to simpler C_n^2 measurement setups. In Supplement 1, we confirm that retraining the models with Π_2 as the sole input feature still yields highly accurate predictions.

The performance of the final Π-ML ensemble is illustrated in more detail in Figs. 4 and 5. The observed (red) and the predicted median evolutions of C_n^2 (black) for the test data are shown in Fig. 4. The evolutions are plotted for the three original sonic heights individually for visualization. The agreement between prediction and observation is high for all levels, although the level-specific ϵ slightly increases with height. For nighttime conditions, the surface layer depth is typically shallower than 10–20 m. Thus, the topmost sonic anemometer at

25 m might be outside the surface layer. In addition, outer layer effects such as wave-induced bursting events can force the turbulence underneath [19]. In such cases, cause (forcing) and effect (turbulence) are vertically separated, so the sonic signal only contains the effect but not the cause. Thus, prediction accuracy decreases without additional upper-air information. Notable errors on all levels mostly occur during atmospheric neutral conditions shortly after sunrise and sunset, where the observed C_n^2 drops as low as 10^{-16}. These drops are overestimated by our ensemble, which is also visible in the 2D correlation histogram of Fig. 5(a) and the quantile–quantile (QQ) plot in 5(b). Panel (a) directly compares observed C_n^2 samples with their ML-estimated counterpart, while panel (b) plots the cumulative density functions of observed and estimated C_n^2 against each other. The overestimation of neutral conditions is visible in both panels as the deviation of the histogram/curve from the ideal 1:1 line (dashed) for $C_n^2 < 10^{-15}$. Simultaneously, the gray 90% confidence band in panel (b) grows, which indicates increasing disagreement between the predictions of the ensemble members. However, less than 8% of C_n^2 measurements are smaller than 10^{-15}, so the regularization of the ML training results in models that favor the center of the C_n^2 distribution, not its tails. Also, the lower signal-to-noise ratio of the sonic anemometers in weak turbulence conditions increases the measurement uncertainty [20]. Since very low turbulence conditions are also not critical for FSOC or astronomy, we argue that little emphasis should be put on these deviations. The regularization mentioned above also explains the minor underestimation visible in panel (b) for observations with $C_n^2 > 10^{-12.5}$, which make up less than 3.5% of the data. Leaving the tails of the distributions aside, both panels of Fig. 5 show excellent performance of our ensemble for most data. Most points in the histogram and the QQ plot in Fig. 5(a) are close to the ideal 1:1 line as quantified by the coefficient of determination of $R^2 = 0.958$ computed on all test data, including the deviating tails. The spread of the correlation distribution around the 1:1 line is symmetric for $C_n^2 > 10^{-15}$. That means the ensemble predictions are well-balanced and not biased toward over or underestimation for most of the C_n^2 range. A brief comparison of Π-ML with two conventional MOST-based C_n^2 parameterizations (W71 [3] and TG92 [21]) in Fig. 5(b) illustrates the potential of improvement by utilizing ML. While W71 and TG92 have the operational advantage of being formulated as analytical equations, they lack the flexibility to capture complex behavior where ML excels. This results in the larger over and underestimations shown in the QQ plots for these popular approaches. Comparing Π-ML to a more traditional ML approach based on [7] (see Section 4 of Supplement 1) also shows significantly higher performance of Π-ML.

In summary, we demonstrated how dimensional analysis constrained with domain knowledge yields non-dimensional surface layer scaling expressions, which enable us to train accurate XGBoost regression models. Our approach has two advantages over C_n^2 parametrizations from the literature. First, the final ensemble produced highly accurate predictions for both daytime and nighttime, while previous models are often limited to one or the other [4]. Second, we expect that the non-dimensional formulation allows making predictions with a pre-trained ensemble for new sonics setup at different heights or locations if the new non-dimensionalized data fall into the original non-dimensional training ranges. The data scaling should enable our ensemble to stay in the interpolation regime longer, i.e., cover a larger dimensional space, compared with traditional ML-based models. At this point, these claims are speculative in nature and need extensive validation. Our final Π-ML ensemble was shown to perform well, regardless of the complex meteorology of Hawai'i [22] and the limited measurement duration of only two months. While the complexity and data sparsity of the MLO campaign limits the applicability of the trained ensemble to other sites, the good performance leads us to posit that our Π-ML methodology might perform well in more favorable setups. Additionally, we observed a strong dependency of C_n^2 on σ_θ^2 (Π_2), suggesting that relatively inexpensive single-level variance measurements might be sufficient for accurate C_n^2 estimation in the surface layer. In conclusion, we presented a powerful, statistically robust physics-informed machine learning methodology (Π-ML) to estimate C_n^2 from turbulence measurements.

Funding. Nederlandse Organisatie voor Wetenschappelijk Onderzoek (P19-13).

Acknowledgments. We are grateful to NCAR for making the MLO C_n^2 data publicly available. FREE project (P19-13) of the TTW-Perspectief research program is partially financed by the Dutch Research Council (NWO).

Disclosures. The authors declare no conflicts of interest.

Data availability. The code implementing the Π-ML methodology is available in Ref. [23].

Supplemental document. See Supplement 1 for supporting content.

REFERENCES

1. H. Hemmati, *Near-Earth Laser Communications* (CRC Press, 2009).
2. H. Kaushal and G. Kaddoum, IEEE Commun. Surv. Tutorials **19**, 57 (2017).
3. J. Wyngaard, Y. Izumi, and S. Collins, J. Opt. Soc. Am. **61**, 1646 (1971).
4. F. G. Smith, *The Infrared & Electro-Optical Systems Handbook: Atmospheric Propagation of Radiation* (Infrared Information and Analysis Center, 1993), Vol. 2 .
5. A. S. Monin and A. M. Obukhov, Cont. Geo. **151**, 1 (1954).
6. M. J. Savage, Agric. For. Meteorol. **149**, 501 (2009).
7. Y. Wang and S. Basu, Opt. Lett. **41**, 2334 (2016).
8. C. Jellen, M. Oakley, C. Nelson, J. Burkhardt, and C. Brownell, Appl. Opt. **60**, 2938 (2021).
9. L. A. Bolbasova, A. A. Andrakhanov, and A. Y. Shikhovtsev, Mon. Not. R. Astron. Soc. **504**, 6008 (2021).
10. C. Su, X. Wu, S. Wu, Q. Yang, Y. Han, C. Qing, T. Luo, and Y. Liu, Mon. Not. R. Astron. Soc. **506**, 3430 (2021).
11. R. B. Stull, *Boundary Layer Meteorology* (Kluwer Academic Publishers, 1988).
12. K. Kashinath, M. Mustafa, and A. Albert, *et al.*, Phil. Trans. R. Soc. A. **379**, 20200093 (2021).
13. S. Oncley and T. Horst, "Calculation of Cn2 for visible light and sound from CSAT3 sonic anemometer measurements," Tech. Rep. (2013).
14. M. Karam and T. Saad, SoftwareX **16**, 100851 (2021).
15. C. Wang, Q. Wu, M. Weimer, and E. Zhu, in *Proceedings of Machine Learning and Systems*, Vol. 3, (2021), pp. 434–447.
16. V. N. Vapnik, *Statistical Learning Theory* (Wiley, 1998).
17. C. Molnar, *Interpretable Machine Learning*, 2nd ed (2022).
18. J. D. Albertson, M. B. Parlange, G. G. Katul, C.-R. Chu, H. Stricker, and S. Tyler, Water Resour. Res. **31**, 969 (1995).
19. L. Mahrt, Annu. Rev. Fluid Mech. **46**, 23 (2014).
20. Ü. Rannik, O. Peltola, and I. Mammarella, Atm. Meas. Tech. **9**, 5163 (2016).
21. V. Thiermann and H. Grassl, Boundary-Layer Meteorol. **58**, 367 (1992).
22. S. Businger, R. McLaren, R. Ogasawara, D. Simons, and R. J. Wainscoat, Bull. Am. Meteorol. Soc. **83**, 858 (2002).
23. M. Pierzyna, "Π-ML: a dimensional analysis-based machine learning parameterization of optical turbulence in the atmospheric surface layer," GitHub (2023) [accessed 17 August 2023], https://github.com/mpierzyna/piml.