Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **509,** 232–245 (2022) https://doi.org/10.1093/mnras/stab2916
Advance Access publication 2021 October 12

Forecasting seeing for the Maunakea observatories with machine learning

Tiziana Cherubini,[1]★ Ryan Lyman[2] and Steven Businger[1]

[1]*Department of Atmospheric Sciences, University of Hawaii, 2525 Correa Road, HIG 350, Honolulu, HI 96822, USA*
[2]*Maunakea Observatory, 640 North A'ohoku Place, #209, Hilo, HI 96720, USA*

Accepted 2021 October 4. Received 2021 September 22; in original form 2021 August 10

ABSTRACT

The staff at the Maunakea Weather Center (MKWC) has provided daily forecasts of optical turbulence for the summit of Maunakea for more than 20 yr. Observational measures of optical turbulence at Maunakea with which to validate official MKWC forecasts have been available since mid-2009. This paper presents a machine-learning approach to translate the MKWC experience into a forecast of the nightly average optical turbulent state of the atmosphere. Maunakea observational and forecast data were collected to build a predictive model of the total and free atmospheric seeing for the following five nights. The motivation for this work is two-fold: to provide a tool/guidance to the MKWC forecaster and allow for a dynamic calibration of the optical turbulence algorithm implemented in the MKWC Weather Research and Forecasting (WRF) model.

Key words: turbulence – atmospheric effects – methods: data analysis – methods: observational – methods: statistical – telescopes.

1 INTRODUCTION

The Maunakea Weather Center (MKWC) has provided forecasts of optical turbulence (OT) for the summit of Maunakea for more than two decades (Businger et al. 2001; http://mkwc.ifa.hawaii.edu). MKWC forecasters have accrued insight into the parameters that impact the forecast through years of daily comparisons between MKWC official forecasts, model guidance, and observational OT measures (Lyman, Cherubini & Businger 2020). The intensity of the turbulent fluctuations of the atmospheric refractive index (i.e. OT) is described by the refractive index structure function, C_n^2 (Coulman 1985). The maximum telescope resolution is defined by a parameter called 'seeing', which is the full width at half-maximum of the long-exposure seeing-limited point spread function of a star image at the focus of a large diameter telescope (Tokovinin 2002). Seeing is the integral of C_n^2 over the light's propagation path and is measured in arc-seconds. Seeing is an astronomical parameter used in this paper and in the MKWC forecasts to summarize the optical turbulent conditions on Maunakea. Seeing and C_n^2 are complex, non-linear functions of a number of atmospheric variables, including potential temperature (θ), temperature, wind, and turbulence (Cherubini & Businger 2013). To model C_n^2 and seeing, and attempt to construct an accurate prediction, requires the implementation of an OT parametrization within a weather model running at very high spatial and temporal resolution (Masciadri, Vernin & Bougeault 1999; Cherubini et al. 2008a; Cherubini, Businger & Lyman 2008b; Cherubini & Businger 2011).

It is very difficult to accurately quantify and predict OT and seeing, and it is even more challenging to pinpoint the temporal variability of OT throughout any given night. Of great benefit to the MKWC forecaster has been the availability of operational measurements of OT and seeing from the multi-aperture scintillation sensor (MASS) and differential image motion monitor (DIMM, MD hereafter; Kornilov et al. 2007) instruments that have routinely operated on Maunakea since the fall of 2009.

As a result of more than a decade of experience in carefully comparing weather patterns and OT measurements, MKWC forecasters learned to recognize the large-scale meteorological variables that foretell future conditions ranging from very good to very bad seeing conditions for the telescopes at the summit of Maunakea, both in terms of weather and OT.

The process of issuing the daily forecast of seeing over Maunakea begins with a review of the latest National Center for Environmental Prediction (NCEP, https://www.emc.ncep.noaa.gov) Global Forecasting System (GFS) analyses and forecasts by the MKWC forecaster. The latest radiosonde profiles of temperature, moisture, and winds with height are also consulted. These profiles measure the atmospheric stability and wind shear. At this early stage, the forecaster has an intuition about the current turbulent state of the atmosphere and for the upcoming nights. This is then refined by consulting higher resolution predictions made by the MKWC's custom implementation of the Weather and Research Forecasting model (WRF; Skamarock et al. 2008), which predicts C_n^2 and seeing (Cherubini & Businger 2013).

The focus of this paper is on understanding the first step of this process and emulating it through machine-learning algorithms applied to archived OT and weather data, which span more than a decade, from 2009 September to present.

Extraordinary progress has been made over the past decade in machine learning (ML hereafter) with consequences that extend to virtually every field, whether educational or commercial. To our knowledge, Milli et al. (2019) were the first study to attempt to nowcast OT by applying ML framework. Their study applies ML to historical observations available at high temporal resolution from Paranal, Chile, to predict the turbulence conditions during the 2 h

★ E-mail: tiziana@hawaii.edu

following the latest observation. The goal of their study is to optimize the telescope's time and instrument scheduling in the very short term.

The goal of this work is to automate the MKWC forecaster's ability to associate average OT observations with large-scale weather patterns and, thereby, anticipate the average OT state of the atmosphere for each of the five nights that the MKWC forecast spans. This study's ambitious expectation is to design a methodology whose performance tend to match the forecaster performance on one hand, while also providing the forecaster with another guidance tool. To accomplish this goal, a suitable data set for ML techniques was first carefully compiled. Then, through the implementation of an unsupervised[1] ML algorithm (*k*-means, https://scikit-learn.org/stable/modules/clustering.html#k-means; Pedegrosa et al. 2011), every past observing night for which OT observations are available is classified as belonging to one of four groups defined by pairing the following: (i) a calm or active OT ground layer, corresponding respectively to low or high values of average ground-layer seeing and (ii) a calm or active free atmosphere, corresponding to low or high values of average free atmospheric seeing.

The effectiveness of this approach is demonstrated through the analysis of the number of observing nights falling in each of the four categories grouped by month. A seasonal trend for the four categories is found that is largely consistent with the MKWC forecasting experience (Lyman et al. 2020).

Finally, a predictive ML algorithm is implemented and trained to learn the underlying associations between nightly averaged OT observations and corresponding weather variables extracted from archived model analyses, and provide a prediction of the expected OT once new weather data are fed into the trained algorithm.

In the operational configuration, for every day the algorithm is run, the following outputs are produced: (i) an estimate of the average total ($\varepsilon_{\rm TOT}$) and free atmospheric ($\varepsilon_{\rm FREE}$) seeing and their standard deviations ($\sigma(\varepsilon_{\rm TOT})$, $\sigma(\varepsilon_{\rm FREE})$) for five upcoming nights and (ii) each of these five nights is classified on the basis of the OT strength in both the ground layer and free atmosphere as belonging to one of the four categories defined by the clustering component of the ML model.

Predictive guidance regarding the average total and free seeing and their projected variability is of clear benefit to the MKWC forecaster. However, the ML approach also opens the way for a dynamical calibration of the C_n^2 algorithm implemented within the MKWC WRF model. The accuracy of the predictive $C_n{}^2$ algorithm depends on a predetermined calibration of the minimum of the turbulent kinetic energy ($TKE_{\rm min}$) (Masciadri & Jabouille, 2001; Cherubini & Businger 2011). The current implementation of the C_n^2 algorithm uses a static profile for $TKE_{\rm min}$ that has a tendency to underestimate the observed total seeing during very good observing conditions and overestimate observed total seeing during strong OT episodes.

A study is ongoing to define a set of four $TKE_{\rm min}$ profiles corresponding to the four categories identified by the ML model. The WRF model will thereafter be run with a $TKE_{\rm min}$ profile, dynamically selected at every model cycle based upon the average OT conditions expected by the ML model. Therefore, the idea to automatize the forecaster intuition has a two-fold purpose: (i) provide another tool/guidance to the MKWC forecaster and (ii) allow for a dynamic calibration of the optical turbulence algorithm implemented in the MKWC WRF model. At time of writing, the ML algorithm presented in this study is run operationally and related products are updated four times daily on the MKWC website.[2]

The paper is structured as follows: Section 2 describes the data used in the study and Section 3 illustrates the building blocks of the implemented ML model and details how it is used to simulate an operational mode. The results of the study are presented in Section 4. Finally, Section 5 draws conclusions and discusses future model refinements.

2 OPTICAL TURBULENCE MEASUREMENTS AND WEATHER MODEL DATA SETS

The observations of optical turbulence used in this study are made by an MD situated at the top of the Canada France Hawaii Telescope (CFHT) instrument tower, approximately 7 m above the ground (see Lyman et al. 2020, Fig. 1). The MD is located between the Gemini and CFHT observatories (Fig. 1), and generally unaffected by terrain or other observatories when winds are from the west and east, which occurs 75 per cent of the time climatologically (Da Silva, 2012). The DIMM measures the integrated optical turbulence through the entire atmosphere, thereby providing an estimate of the total atmospheric seeing ($\varepsilon_{\rm TOT}$). The MASS does not sense optical turbulence near the ground but reconstructs turbulence profiles at six altitudes ($h = 0.5$, 1, 2, 4, 8, 16 km) above the telescope (Tokovinin & Kornilov 2007) and provides an estimate of the free atmospheric seeing ($\varepsilon_{\rm FREE}$). Ground-layer seeing, ε_G, is calculated using the following formula (Skidmore et al. 2009):

$$\varepsilon_g = \frac{\varepsilon_{\rm TOT} - \varepsilon_{\rm FREE}}{|\varepsilon_{\rm TOT} - \varepsilon_{\rm FREE}|} \left|\varepsilon_{\rm TOT}^{5/3} - \varepsilon_{\rm FREE}^{5/3}\right|^{3/5}. \quad (1)$$

The MD instruments are set to operate under the following thresholds: relative humidity <85 per cent and wind speed $<14\,{\rm m\,s^{-1}}$ ($\sim 50\,{\rm km\,h^{-1}}$). At time of writing, the MD has operated for over 2800 nights since 2009 September (real-time data are available at http://mkwc.ifa.hawaii.edu/current/seeing/).

The data set used in this study is an extended version of a *seeing catalogue* available online at http://mkwc.ifa.hawaii.edu/current/seeing/analysis/NbN/?file=mkseeing. This data set was constructed to allow the MKWC forecaster to more easily correlate optical turbulence with large-scale (synoptic) weather patterns. In addition to the MD observations, the seeing catalogue includes carefully selected variables available in the GFS operational analyses (https://rda.ucar.edu/datasets/ds084.1, National Centers for Environmental Prediction/National Weather Service/NOAA/US Department of Commerce, 2015). GFS analyses are a blend of observations and a model first guess field from the previous GFS run. The data set used in this study, hereafter referred to as the ML seeing catalogue, is a tabular data set for each night (UTC date) containing the variables listed in Table 1. These variables are extracted from the GFS analyses validating 12 UTC (2am HST, the middle of the observing night on Maunakea) for the corresponding catalogue date. The ML seeing catalogue also includes the following user derived variables: (i) shear, $|\Delta U/\Delta z|$, between 600 and 200mb; (ii) ground-layer seeing, ε_G; and (iii) status, which is a categorical variable. The status indicates the probability the corresponding night is a good (1) or bad (0) night in terms of weather (high winds, cloud cover, fog, precipitation, etc.), and indicates whether astronomical observations

[1]Unsupervised learning is a type of algorithm that learns patterns from untagged, unclassified data, in contrast to supervised learning where data are tagged by a human. Unsupervised learning exhibits self-organization that captures patterns as neuronal predilections or probability densities.

[2]MKWC Machine Learning prediction (http://mkwc.ifa.hawaii.edu/forecast/mko/trends/index.cgi?trend=latest&param=keyvar&model=neural&res=).

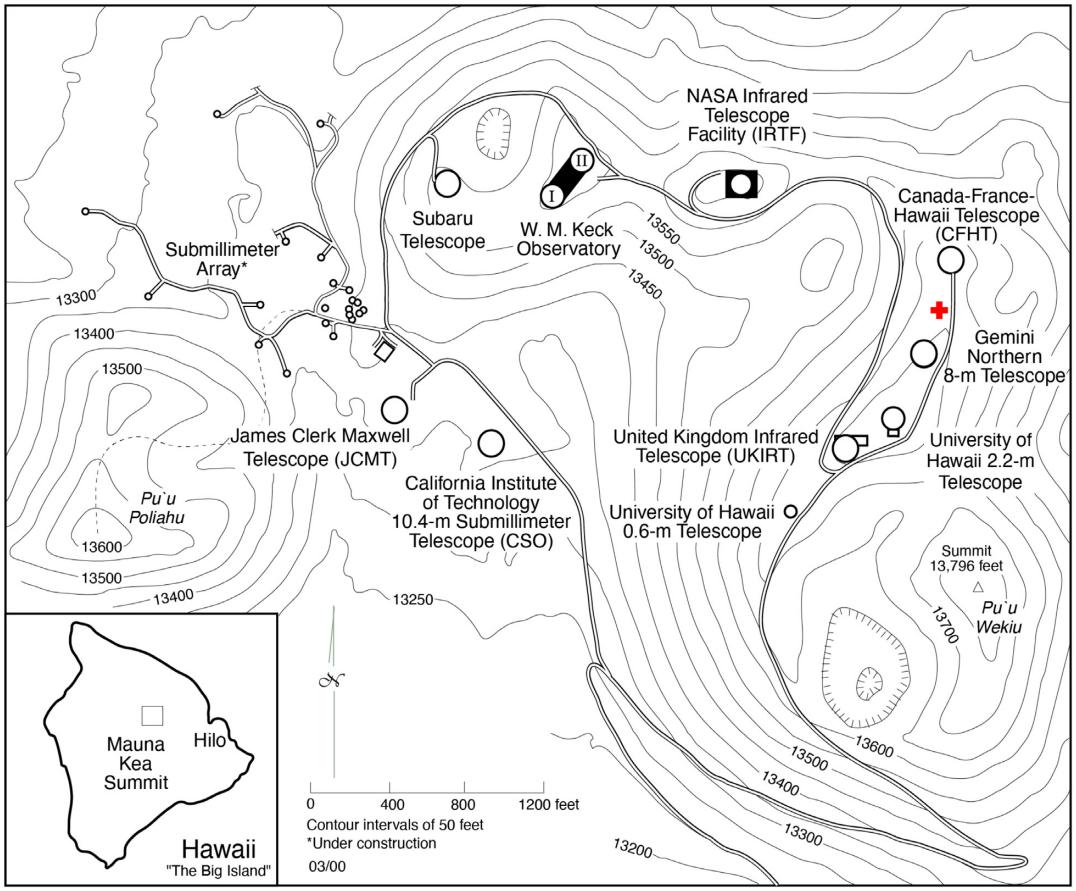

Figure 1. Detailed map of the summit of Maunakea. The red cross indicates the position of the MD between the Gemini and CFHT observatories.

Table 1. Variables included in the tabular ML seeing catalogue.

MD optical turbulent data (five targets)	GFS variables (33 weather features)
(i) Nightly averaged DIMM seeing ($\varepsilon_{\rm TOT}$)	(i) Potential temperature (θ) on all GFS levels between 650 and 200 mb by 50 mb intervals
(ii) Nightly averaged DIMM seeing standard deviation	(ii) Wind speed on all GFS levels between 650 and 200 mb by 50 mb intervals
(iii) Nightly averaged DIMM number of measurements (not a target)	(iii) Wind direction on all GFS levels between 650 and 200 mb by 50 mb intervals
(iv) Nightly averaged MASS seeing ($\varepsilon_{\rm FREE}$)	(iv) 600 mb relative humidity
(v) Nightly averaged MASS seeing standard deviation	(v) 600 mb precipitable water (PW)
(vi) MASS number of measurements (not a target)	(vi) Vertical shear of the horizonal wind between 600 and 200 mb
(vii) Status (0,1)	

are possible or not. Because the MD instrumentation does not record the specific conditions for observations (active observation, closure due to weather, or closure due to maintenance) the status variable has been engineered as follows: status = 0 when MD data are not recorded and $RH > 59\ per\ cent$ or $w_{600mb} \geq 13\ m\,s^{-1}$; or status = 1. The thresholds for RH and wind speed are conservative thresholds chosen on the basis of the forecaster's experience and recognized as indicators of bad/potentially bad weather nights. These constrains might be too conservative and result in the exclusion of few transient weather days that potentially could have been active observing days. The data set spans from late 2009 September, when MD started operating, to the end of 2020.

3 METHODOLOGY

This section describes the components of the implemented MKWC ML model (Fig. 2) which aims to: (i) predict the nightly average optical turbulent state of the atmosphere in terms of predicted average total and free atmospheric seeing and their anticipated standard deviations (right block in Fig. 2) and (ii) classify the strength of the optical turbulence in both the ground layer and free atmosphere (left block in Fig. 2).

Before using the ML *seeing catalogue* in our ML algorithm, the data set is pre-processed as follows: (i) only nights with more than 50 MD OT measurements per night are retained and (ii) model weather data are checked for spurious values, such as null relative humidity values or spurious potential temperature values (possibly due to conversion/storage/digitalization errors), and these repaired appropriately. The null relative humidity is substituted with 1 per cent values; the spurious potential temperature is filled in with adjacent values.

From this point onwards, the authors will refer to input features (Fig. 3) as the atmospheric variables used as input in the ML model and targets (Fig. 4) the OT variables that are the targeted outputs

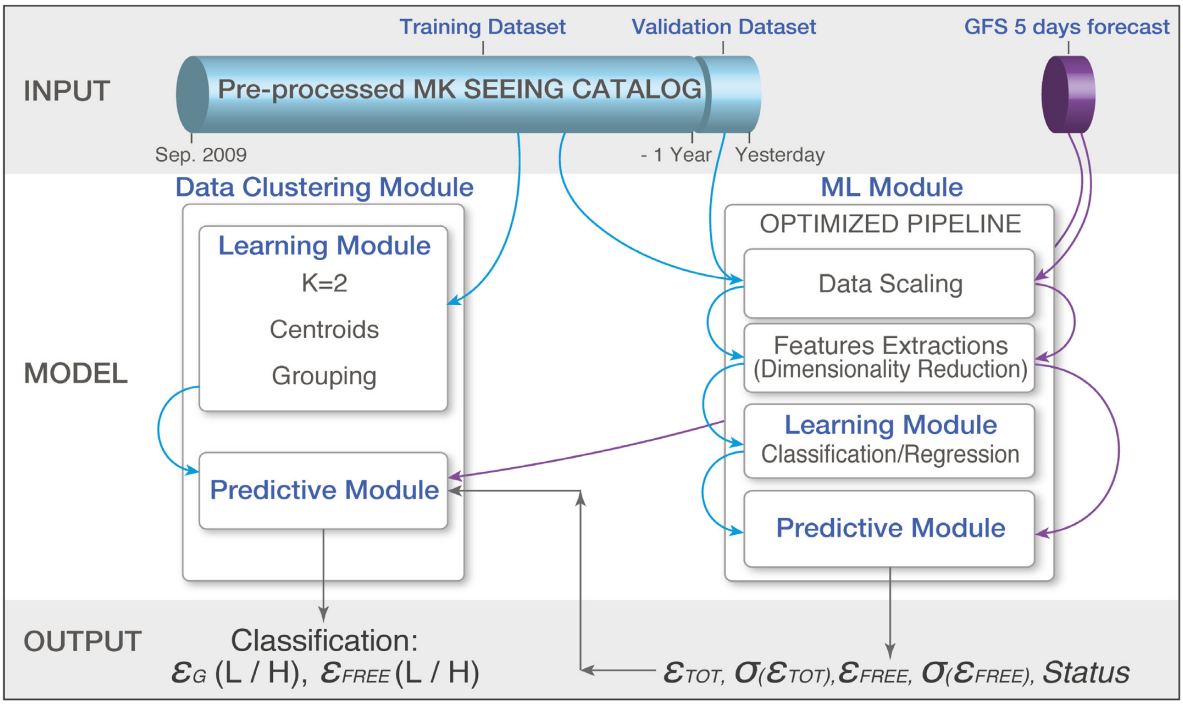

Figure 2. Block diagram of the implemented operational ML model.

for the ML model. The input features follow either a Gaussian distribution (θ) or a lognormal/Poisson distribution (wind speed, pw, RH), and a bimodal distribution for wind direction, indicating an easterly and westerly prevalence, as expected.

3.1 Data clustering

The MKWC forecaster's main indication for the classification of OT strength in the ground and free atmospheric layer comes from the GFS wind analysis at 600 and 250 mb. Optical turbulence is driven by many factors but the wind speed at summit level and at the 250 mb reference level have proven, over time, to be dominant factors to first order, for estimating the average optical turbulent state of the ground layer and free atmosphere, respectively (Lyman et al. 2020). The analysis of a year of data shows a correlation coefficient $\rho \sim 0.35$ between wind speed at summit surface and total seeing (Fig. 5a). Chun et al. (2009, fig. 10) also discuss the relationship between surface winds and ground layer seeing. Similarly, although perhaps less expected, a correlation coefficient $\rho \sim 0.32$ between free atmospheric seeing and the GFS 250 mb wind speed (Fig. 5b) is found, which is a manifestation of the upper level jet stream's strength and proximity.

To translate the MKWC forecaster's associative skill into an ML algorithm, a clustering algorithm for classification is chosen. The *k*-mean clustering algorithm used in this study is an unsupervised ML method, largely used in data mining, that aims to find groups of observations, or clusters, that share similar characteristics (Pedregosa et al. 2011).

In this application, the *k*-mean clustering algorithm classifies: (i) average optical turbulent behaviour of the ground layer for each recorded night into good (low) seeing (L_G) or poor (high) seeing (H_G) ground layers and (ii) average optical turbulent behaviour of the free atmosphere for each recorded night into good (low) seeing (L_{FREE}) and poor (high) seeing (H_{FREE}) free atmospheres.

Two separate instances of the *k*-mean algorithm are run with $k = 2$: one targets ground-layer seeing and the second one targets free atmospheric seeing. For the ground layer, the ML algorithm uses as input variables the derived ground-layer seeing, ε_G, measured total seeing and its standard deviation as measured by DIMM, along with the wind velocity at 600 and 250 mb levels, and shear derived from the GFS analyses. All variables are nightly averages.

For the free atmosphere, the ML algorithm uses as input the observed free seeing and its standard deviation as measured by MASS, wind velocity at 600 and 250 mb level, and shear from the GFS analyses; all variables are nightly averages.

The input variables for the ML clustering algorithm are normalized to share a common scale. This pre-processing step is particularly important in cluster analysis because groups are defined based on the distance between points in mathematical space, while the chosen variables have different units and therefore span on different scales. The choice of the selected variables is driven by the relational insight of the MKWC forecaster derived from experience.

The ML procedure runs the two *k*-mean instances: a cut-off date in the timeseries underlying the ML seeing catalogue is selected and data spanning from the beginning of the data base to the selected date are used by the *k*-mean algorithm to cluster ground-layer seeing and free atmospheric seeing in two groups, and to find the relative centroids. The trained ML algorithm can then classify, through the *k*-mean prediction method, any night not included in the training data set as belonging to the L_G or H_G cluster of ground-layer seeing or belonging to the L_{FREE} or H_{FREE} cluster of free atmospheric seeing. Any given night not in the training data set can further be classified as belonging to one of the four groups defined by all the possible combinations of the two pairs of L/H classes.

3.2 Machine learning predictive model

This section describes how the ML algorithm is implemented, with the following objectives: (i) learn the associations between MD

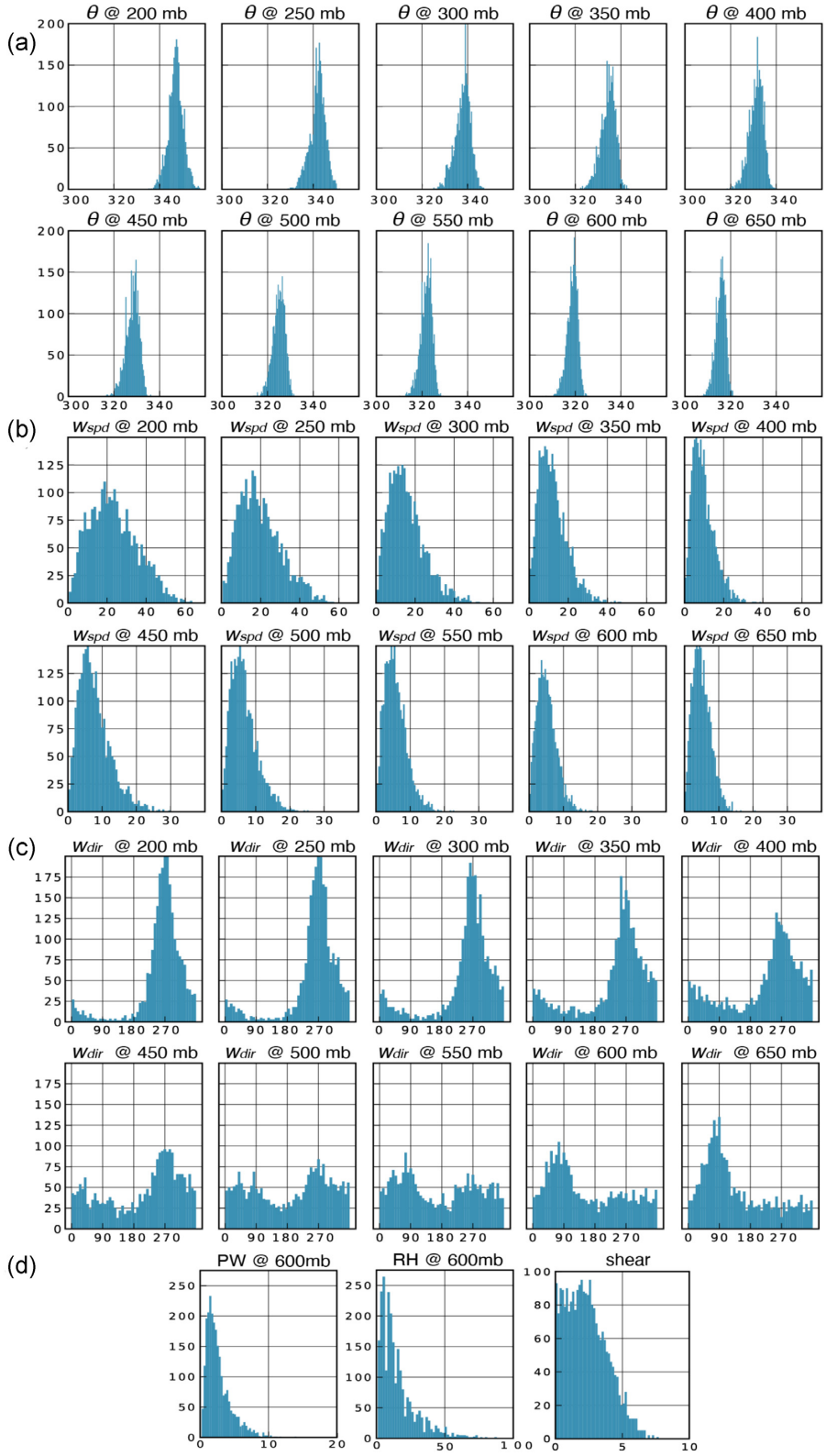

Figure 3. Input features (33) to the ML model, extracted from the ML seeing catalogue: (a) potential temperature θ; (b) wind speed; and (c) wind direction on every level from 650 to 250 mb by 50 mb; (d) precipitable water, relative humidity at 600 mb and shear between 600 and 250 mb.

average measurements and selected average GFS variables when a training data set is provided; (ii) anticipate through a classification (logistic regression) algorithm the probability (between 0 and 1) that a given night is going to be a good (1) or bad (0) observing night; and (iii) predict through a regressor algorithm the average OT behaviour in both the ground layer and free atmosphere, once weather data from a model analyses (for case studies) or forecasts (operational applications) are available.

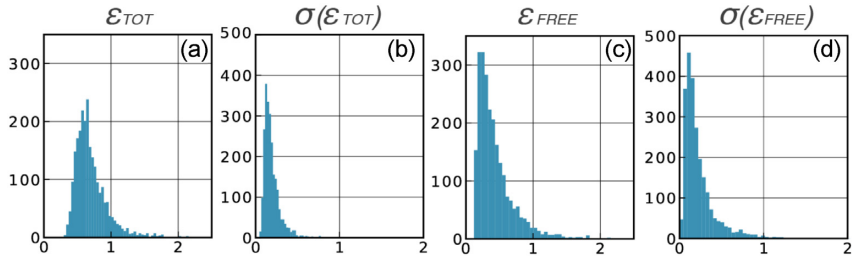

Figure 4. Targets in the ML model: nightly averaged total (a) and free atmospheric (c) seeing as measured by MD, and their standard deviations (b and d, respectively).

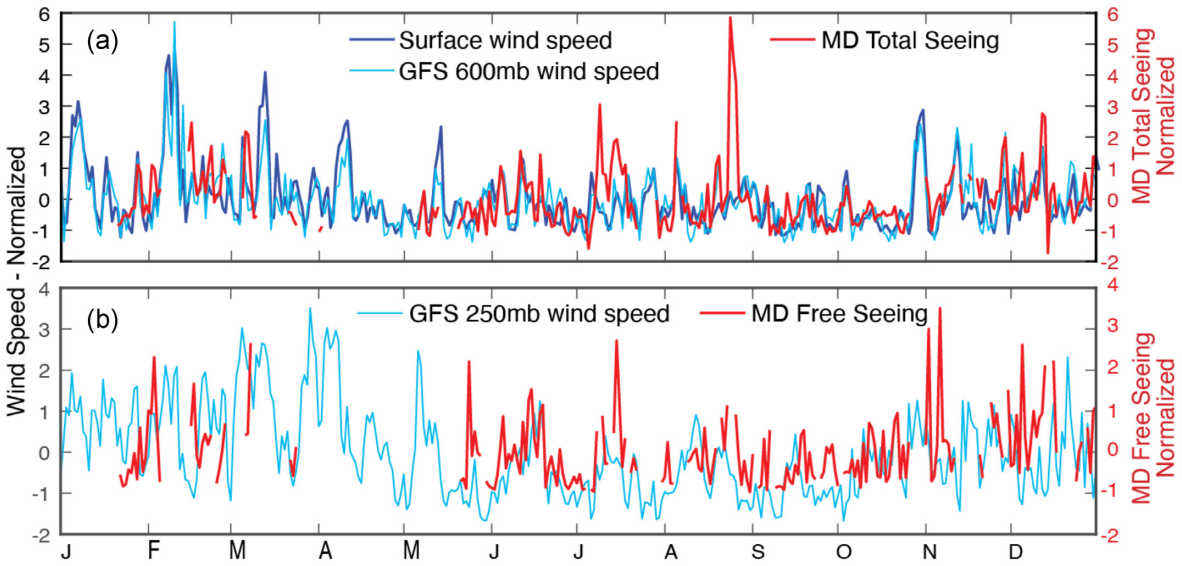

Figure 5. (a) Timeseries of normalized surface wind speed as recorded by the surface anemometer (blue solid line) located at the top of the CFHT instrumented tower, normalized GFS 600 mb wind speed (light blue solid line), and normalized total seeing (red solid line) as recorded by the MD. The timeseries span over the full year 2020. Data normalization allows the data to be compared on similar scales.

The compiled ML seeing catalogue (Table 1) is the input data set for our ML model: the weather variables are the model's features, while the OT variables and status are the model targets. Once the cut-off date in the ML seeing-catalogue timeseries is selected, data from the beginning of the data base to a year before the selected cut-off date are used as the training data set, D_T. The year of data prior to the cut-off date is used as the validation data set, D_V. A prediction can then be made on the remaining dates from the cut-off date to the most current time in the timeseries, the testing data set, D_t.

Both the regression and classification components of the ML model are optimized through the following sequence: (i) data scaling; (ii) dimensionality reduction; and (iii) regression/classification. A grids-search cross-validation[3] process is conducted over this pipelined sequence to tune the involved hyperparameters and select those that minimize prediction error on the validation data set. Various regressors were tested in the ML algorithm's workflow and the Bayesian Ridge regressor[4] was found to produce the best scores.

A simple logistic regression module was chosen for the categorical portion of the ML model to predict the target status. The ML model is run once to learn, validate, and predict the categorical variable status, and run again to learn about, validate, and predict the main four target variables: average total seeing and its standard deviation; free atmospheric seeing and its standard deviation. While a more correct approach to this problem would be to implement a multi-output regression model, in this study the prediction of the various targets is treated as one separate problem for each target variable. This assumes that the target outputs are independent of each other, which may be incorrect. Nevertheless, this approach provides an effective prediction and the results are very promising. Therefore, the ML algorithm as currently constructed is worth exploring as a performance baseline, with the possibility of improving or refining the ML algorithm in future.

3.3 Operational implementation

This subsection describes how the performance of the implemented ML model is then tested for a yearlong timeframe, which simulates an operational set-up. The testing period spans from 2020 January 1 to December 31. The generic date in this timeframe is referred to as the current date and the day before the current date is referred to as the cut-off date. The cut-off date is therefore a date in the ML seeing catalogue that is in the past. For each current date, and each synoptic cycle (00, 06, 12, 18 UTC), the GFS forecasts validating every 12-UTC time (2 AM HST) up to 5 d from the current one,

[3] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html.
[4] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.BayesianRidge.html.

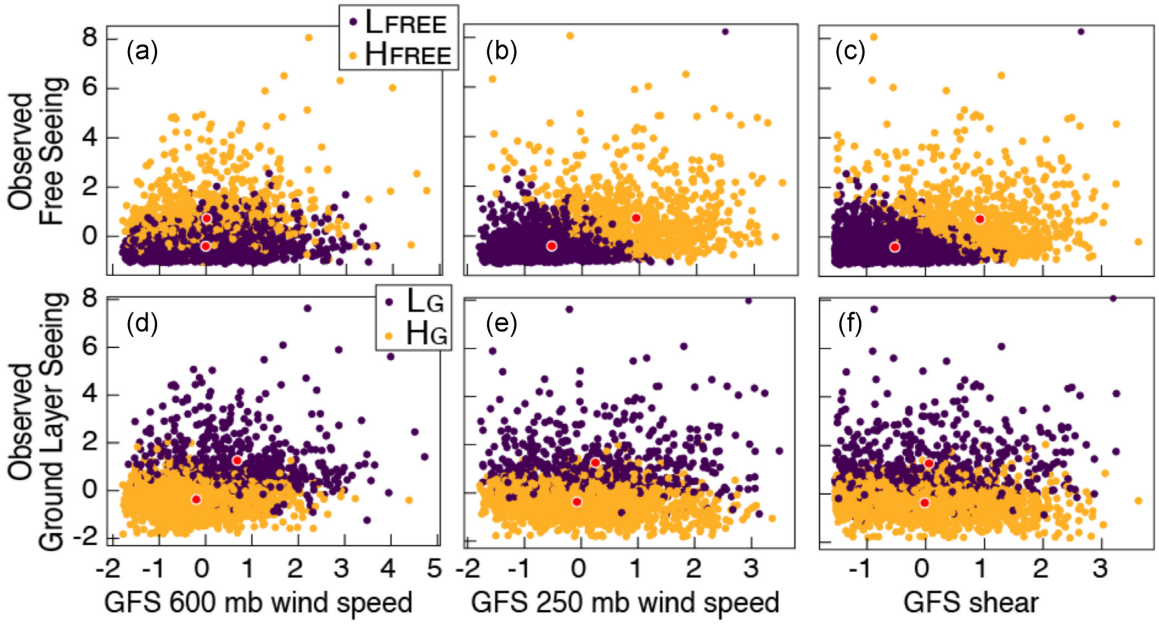

Figure 6. Clusters as result of running the *k*-mean algorithm on the full data set (2009–2020) for both the free-atmosphere seeing (panels a, b, c) and the ground-layer seeing (panels d, e, f). Clusters are four-dimensional but are plotted in two-dimensions, with normalized scales, as a result of considering one feature at a time. Axes are scaled. Centroids are shown as red dots.

are used to extract and derive the environmental variables defining the average weather conditions at and above the Maunakea summit in ML seeing catalogue form. Once a date is selected as current: (i) the portion of the ML seeing catalogue from its beginning (2009 September) to a year before the cut-off date is used as training data set for the ML algorithm; (ii) the yearlong timeframe in the ML seeing catalogue up to the cut-off date is used as the validation data set for the ML algorithm; and (iii) the environmental variables from the GFS forecasts (therefore data belonging in the simulated future) are used as input features. The desired outputs (targets) are the predictions of the average total and free atmospheric seeing and their standard deviations for each synoptic cycle and the following five nights.

Once the ML predictions on the average OT state of the atmosphere for the next five nights are available, a classification of these five nights is then possible via the ML *k*-mean algorithm, as described in Section 3.1. The ML *k*-mean algorithm uses as a testing data set the input features extracted by the GFS forecasts valid for the next five nights. The full output, once both the ML and *k*-mean portion of the model are run, is composed of the OT predictions, their standard deviations, and the associated L/H classes for each night (Fig. 2).

4 RESULTS

4.1 Data clustering

A distinct separation of classes and centroids are evident in Figs 6(b) and (d), consistent with the known large influence of the wind speed at 250 mb on free atmospheric seeing and the wind speed at 600 mb on ground-layer seeing, respectively (see Tables 2 and 3). Given the complex nature of OT and the low number of clusters requested ($k = 2$) it was not expected that fully separated clusters would be obtained in all cases; the resulting separation seen in Fig. 6 is very encouraging.

Table 3 shows that the ML *k*-mean for the classification of the ground layer OT provides a centroid for which the mean total seeing

Table 2. L_{FREE} and H_{FREE} classes in terms of the mean OT parameters as defined by the ML *k*-mean algorithm when characterizing the free atmosphere OT behaviour.

	ε_{TOT}	$\sigma(\varepsilon_{TOT})$	ε_{FREE}	$\sigma(\varepsilon_{FREE})$	ε_G
L_{FREE}	0.65	0.18	0.32	0.15	0.52
H_{FREE}	0.78	0.22	0.64	0.35	0.60

Table 3. L_G and H_G classes in terms of mean OT parameters as defined by the ML *k*-mean algorithm when characterizing the ground layer OT behaviour.

	ε_{TOT}	$\sigma(\varepsilon_{TOT})$	ε_{FREE}	$\sigma(\varepsilon_{FREE})$	ε_G
H_G	1.03	0.32	0.68	0.34	0.92
L_G	0.61	0.16	0.37	0.19	0.44

for the L_G classes is 0.61 arcsec, which is very close to Maunakea's observed median seeing of 0.65 arcsec (Lyman et al. 2020), while the centroid for the H_G class provides a mean total seeing value close to 1 arcsec. Similarly, the L_{FREE} class for the characterization of the free atmosphere, shows a centroid for which the mean free atmospheric seeing corresponds to ∼0.32 arcsec during calm nights, which is very close to the observed median free seeing of 0.35 arcsec (Lyman et al. 2020), while the H_{FREE} class identifies OT episodes characterized by free atmospheric seeing values about double in strength (∼0.6 arcsec).

Regarding the ground layer, the number of calm nights dominates the L_G events, which climatologically peak during summer, with maxima in June and September/October, which are historically the months with best seeing quality (Fig. 7b). The occurrence of more active nights and poorer seeing is greater during the winter season. During the middle of the summer, there are fewer episodes of high surface winds and/or strong winds aloft that account for the smaller maximum in poor seeing.

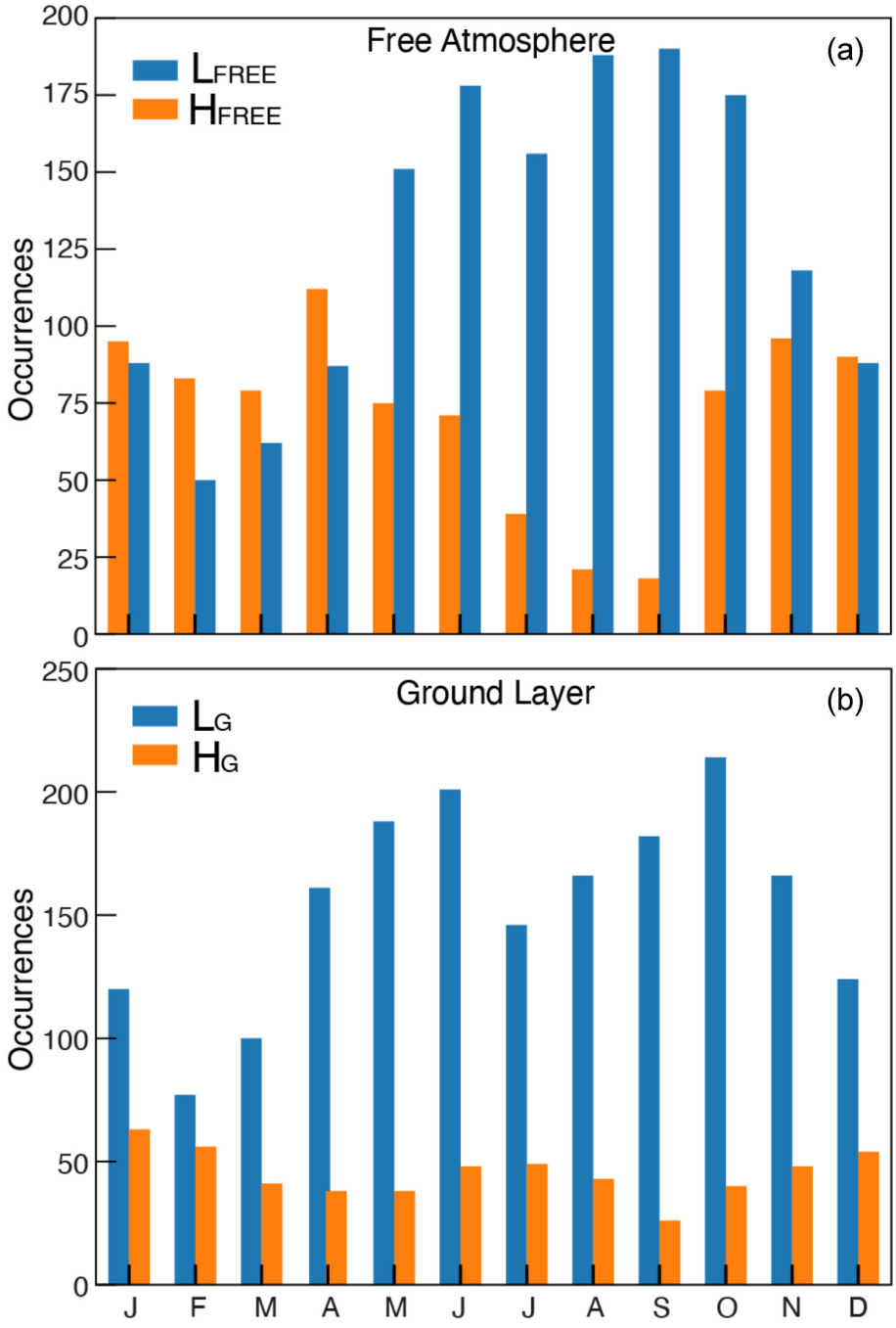

Figure 7. Number of cases from the full catalogue (2009–2020) separated by classes L/H, for both the free atmosphere (a) and the ground layer (b).

Regarding the free atmosphere, the number of less active nights with better seeing peaks during summer as expected (Fig. 7a). During the winter season, the westerly jet stream aloft increases in strength and moves closer to Hawaii, resulting in more weather disturbances entering the subtropical Central Pacific area from the north, which increases the chance for turbulence aloft. These findings, which were obtained by applying a clustering technique on the ML seeing catalogue, are consistent with the seeing climatology observed over the years (Lyman et al. 2020).

Finally, the pairing of low OT events in both the ground layer and free atmosphere (L_G, L_{FREE}) is the largest of the four possible combinations (Fig. 8), clearly peaking during the summer months. This result is consistent with the fact that Maunakea is considered the best ground site on Earth for optical astronomy. Events during which both the ground and free atmospheric seeing are poor are more likely to occur during winter when the polar jet-stream is climatologically closest to Hawaii. It is interesting that during spring and fall the chance of having good seeing in the ground layer yet poor seeing aloft is greatest. This may reflect the fact that during these seasons the circumpolar jet stream tends to exhibit higher Rossby-wave numbers and the flow aloft has greater meridional excursions that impact the subtropics.

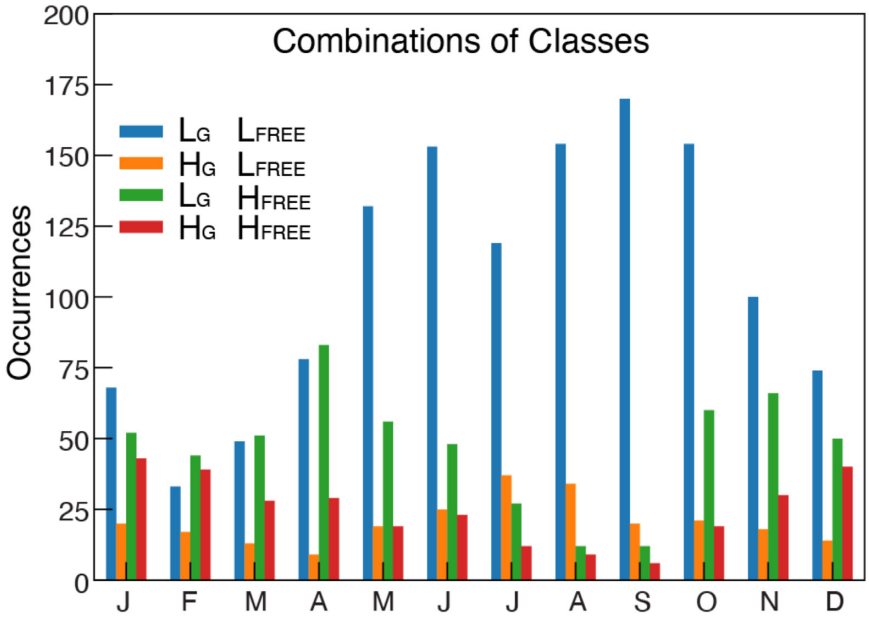

Figure 8. Number of cases from the full catalogue (2009–2020) separated by the combination of classes L/H, as defined by the ML *k*-mean algorithm. Blue bars refer to cases for which both ground layer and the free atmosphere seeing is good (L_G, L_{FREE}); orange bars refer to cases for which ground-layer seeing is poor (H_G) and the free atmosphere seeing is good (L_{FREE}); green bars refer to cases for which ground-layer seeing is good (L_G) and the free atmosphere seeing is poor (H_{FREE}); red bars refer to cases for which both the ground layer and the free atmosphere contribute to poor seeing (H_G, H_{FREE}).

Table 4. ML data sets in the experiment described in Section 4.2.

Data set	Timeframe
D_T	2009 to 19 April 2019
D_V	20 April 2019 to 19 April 2020
D_t	20 April 2020 to 31 December 2020

4.2 Training and validation of ML predictive model

With reference to Section 3.2, we select the cut-off date of 2020 April 20 as an explanatory example. The corresponding timeframes for D_T, D_V, and D_t are listed in Table 4. One of the main steps of the implemented ML module, after data scaling, is dimensionality reduction by selection of the most important input features in the data sets, that is, those providing better scores in the underlying cross-validation analysis.

The number of input features selected and passed to the final step of the ML algorithm, the regressor, varies depending on the chosen date and the size of the training data set. The training process identifies between 14 and 18 most important features, depending on the cut-off date in the timeseries, when the target is the total observed seeing, and between 22 and 25 most important features when target is the free atmospheric seeing.

The dimensionality reduction process reveals as most important[5] those variables that the forecaster also recognizes through experience as dominant factors in predicting seeing. The wind speed on the model levels closest to the summit, that is 650, 600, and 550 mb, are the most important variables in determining the total atmospheric seeing, followed by potential temperatures at 650 mb and, with minor contributions, potential temperature from summit level up to the middle of the troposphere. Wind direction, shear, and moist variables (RH and PW) contribute little by comparison (Fig. 9a).

For observed average free atmospheric seeing the analysis is, as expected, more complex. Potential temperature largely dominates in the middle atmosphere. The impact of wind peaks at 250 mb (jet level) and falls off gradually at lower altitudes, except for a secondary maximum at 650 mb (just below summit level). The calculated wind shear provides redundant information to that in the wind profile. As expected, moisture related variables contribute very little (Fig. 9b). These results are consistent with the experience of the MKWC forecaster (Lyman et al. 2020).

The ML algorithm shows good predictive skill, particularly in terms of the total atmospheric seeing, with a correlation coefficient of 0.67 (Fig. 10a). As expected, the free atmospheric seeing prediction is slightly less accurate, with a correlation coefficient of 0.47. A few outliers are seen, more numerous for free seeing prediction, where the ML algorithm tends to underestimate stronger OT episodes. A possible explanation for this observation is that the stronger turbulence events are the least numerous in the seeing catalogue (Fig. 4b), therefore providing a smaller showcase for training. It is reasonable to assume that the ML algorithm will improve performance in predicting strong OT events as their number increases with time.

Capturing the behaviour of OT in the free atmosphere is more difficult than capturing the behaviour of the total seeing (compare Fig. 10a and c), which may be due to a greater underlying complexity for the regressor to handle. Also, Masciadri, Lombardi & Lascaux (2014) pointed out that MASS data could potentially carry biases when compared to higher resolution OT vertical profilers like the G-Scidar. These biases could affect any model and/or calibration based on them. Nevertheless, at the time of this writing, the MD is the only consistent OT data source available to the MKWC and, although potentially biased, these seeing observations remain a critical tool in the quest to better understand and predict OT on Maunakea and develop predictive tools like the one presented in this study.

[5]The input features' scores of the dimensionality reduction method, SelectKbest, used in the modelled pipeline is often referred to as 'features' importance and it is unitless.

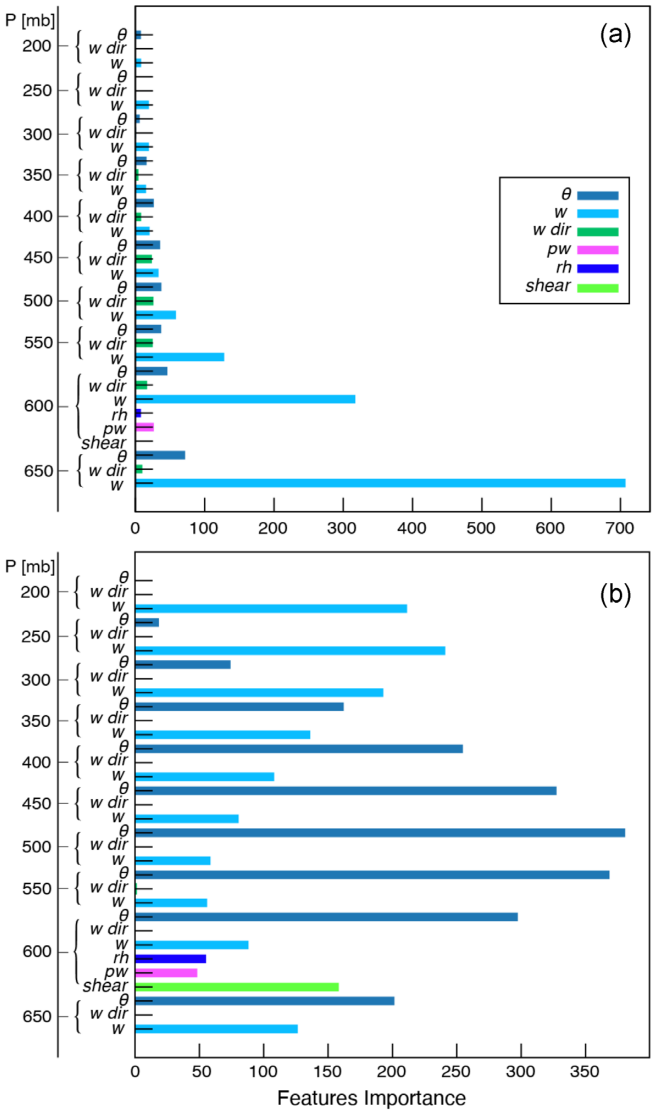

Figure 9. Features importance when observed total (a) and free atmospheric (b) seeing are targets for the ML model. Light blue bars indicate wind speed, green bars indicate wind direction, and navy blue bars indicate potential temperature. Wind shear between 600 and 250 mb (light green bar), relative, relative humidity (blue bar), and total precipitable water (magenta bars) at summit level are also shown.

4.3 Operational set-up: a yearlong experiment

This section summarizes the results of the experiment described in Section 3.3, which simulates an operational set-up and a yearlong timeframe. For each date in 2020, the ML algorithm was run and the GFS forecasts from the four synoptic cycles are used to provide input features validating at 12 UTC (2 AM HST) for five nights past the chosen date. For each day of the year a prediction, stratified by the model forecast hour validating the 12 UTC, is made. For example, the 00 UTC GFS cycle provides the input features for the 12, 36, 60, 84, and 108 h of forecast, the 06 UTC GFS cycle provides the input features for the 6, 30, 54, 78, and 102 h of forecast, and so on for the remaining 12 and 18 UTC cycles. The ML algorithm predicts the corresponding OT variables.

The overall result is a multiprediction for each night, stratified by forecast hour. The figures discussed in the next paragraph only refer to the 12 h prediction. Similar plots have been made and analysed for all other forecast time-steps but are not shown here as they convey very similar graphical contents: the slow increase of the GFS forecast error with the forecast time results in small differences in the ML algorithm's performance as forecast time increases (Table 5).

Overall the predicted behaviour of the total and free atmosphere OT (Fig. 11) is well captured by the ML algorithm and seasonality is clearly reproduced. It is also reassuring to see that the predicted standard deviation tends to be smaller in correspondence to lower values of predicted seeing, which is also the case for the corresponding measurements. Fig. 12 shows the scatterplots and corresponding statistics for this experiment, which are consistent with the results in Fig. 10.

As discussed in previous section, episodes of strong observed OT degrade some of the statistics. The differences between the results shown in Figs 10 and 12 are due to differences in the data source from which the weather variables are inferred. In order to test the ML algorithm under optimal conditions, GFS analyses are used in the former experiment. These forecasts are also used in the latter experiment to test the ML model in an operational setting, which is when a prediction on the future state of the atmosphere is desired and weather analyses of the atmosphere for that time are not yet available. The operational configuration carries the GFS forecast error into the ML algorithm. Therefore, a slight degradation of the statistical scores for this set-up is to be expected.

Similarly, there is a degradation of the statistical measures as the forecast hour increases, corresponding to an expected increase in the GFS forecast error with increasing forecast time (Table 5). The statistics show a good correlation coefficient (∼0.6) and good bias corrected rms (∼0.2) for the predicted total seeing for the forecast hours 6, 12, and 18, which refer to the first observing night.

Once the ML algorithm is run for 2020, each day is classified in terms of the L/H classes for the OT turbulence in the ground layer and free atmospheric layer. This will potentially have a large impact on a follow-up study that will focus on a dynamic calibration of the OT algorithm implemented within the MKWC WRF model. The a priori knowledge of the OT strength in both the ground and free atmosphere layers will be used to select the appropriate calibration for the $TKE_{\rm min}$ before a WRF simulation is made. The results from this study will be the main focus of a follow-up paper.

5 SUMMARY, CONCLUSIONS, AND DISCUSSION

By carefully comparing large-scale meteorological variables relevant to Maunakea with operational measurements of OT from the MD instruments and forecast output from NOAA's GFS weather model, the MKWC forecaster learned to anticipate the average turbulent state at the summit of Maunakea and in the free atmosphere above for the five upcoming nights. This paper leverages ML algorithms to construct an automated prediction model that mimics the MKWC forecaster's ability to associate average OT observations with large-scale weather patterns seen in the GFS model output, and thereby anticipate the average OT state of the atmosphere for each of the five nights that the MKWC forecast spans. The results of this study show that it is possible to capture the insight of the Maunakea forecaster in predicting the average OT behaviour at Maunakea (MK) and produce a novel tool that the MK forecaster and the MK telescope operators can consult to help make forecast decisions. In addition, the success for this approach opens a new door that allows for the development of a dynamic calibration of the optical turbulence algorithm that estimates OT as part of the WRF model cycle.

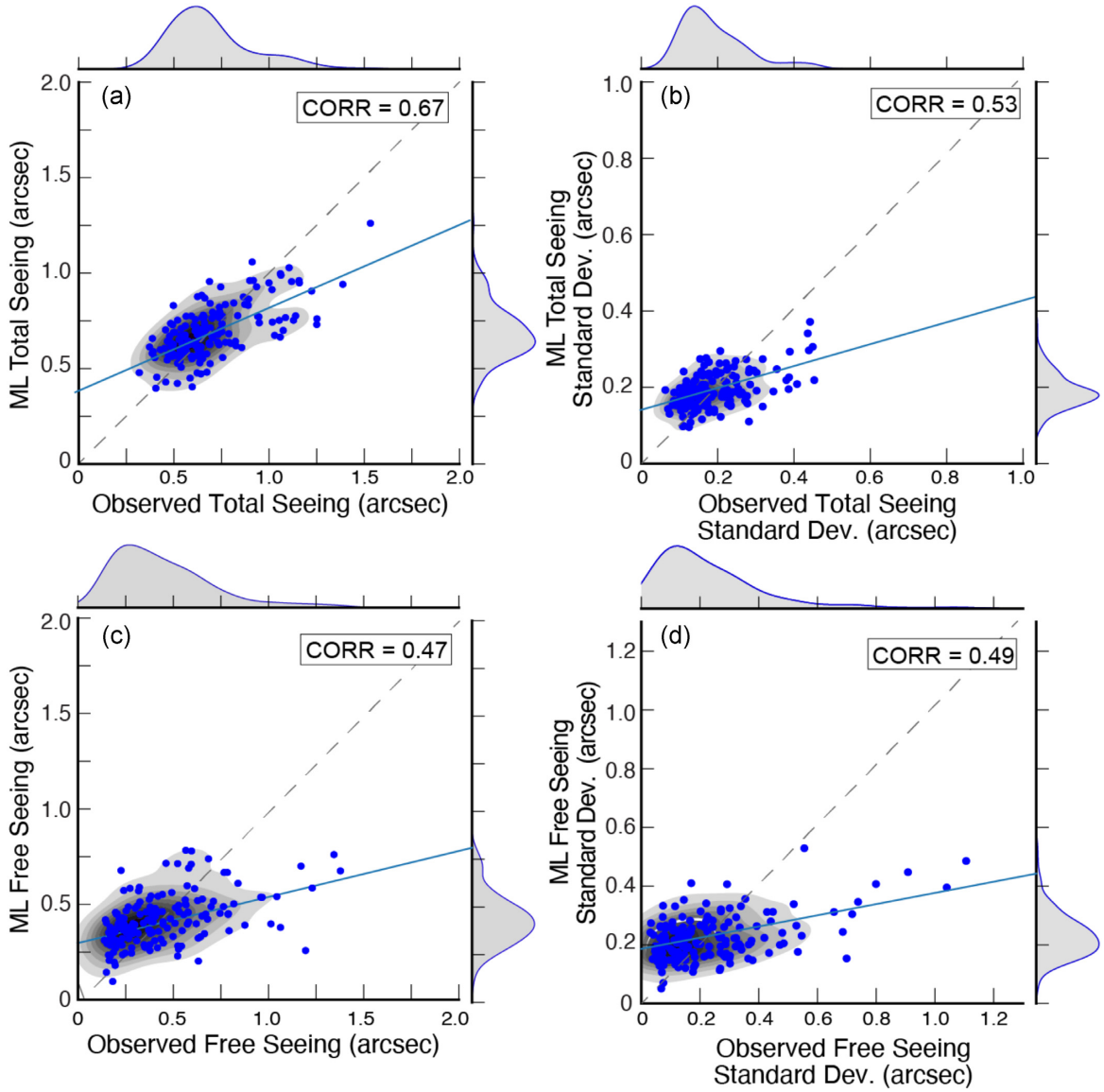

Figure 10. Scatterplots of the ML predicted versus observed: (a) total seeing and its (b) standard deviation; (c) free atmospheric seeing and its (d) standard deviation for the testing data set Dt. Correlation coefficients are also shown.

Table 5. Statistics for the ML predicted total and free atmospheric seeing and their standard deviations, at various forecast hours, for the yearlong 2020 experiment. The highlighted grey row refers to the 12th hour forecast referred to in Fig. 12 and discussed in the text.

Forecast hour	BIAS ε_{TOT}	RMSE ε_{TOT}	CORR ε_{TOT}	BIAS ε_{FREE}	RMSE ε_{FREE}	CORR ε_{FREE}	BIAS $\sigma(\varepsilon_{TOT})$	RMSE $\sigma(\varepsilon_{TOT})$	CORR $\sigma(\varepsilon_{TOT})$	BIAS $\sigma(\varepsilon_{FREE})$	RMSE $\sigma(\varepsilon_{FREE})$	CORR $\sigma(\varepsilon_{FREE})$
6	−0.002	0.199	0.639	−0.035	0.247	0.491	0.002	0.086	0.476	−0.009	0.200	0.424
12	−0.006	0.204	0.612	−0.040	0.246	0.493	0.001	0.088	0.449	−0.010	0.201	0.413
18	−0.009	0.207	0.599	−0.040	0.250	0.472	0.001	0.087	0.462	−0.010	0.203	0.383
24	−0.007	0.210	0.584	−0.040	0.250	0.474	0.002	0.087	0.449	−0.009	0.205	0.359
36	−0.009	0.212	0.568	−0.045	0.256	0.438	0.002	0.087	0.460	−0.011	0.206	0.356
48	−0.010	0.211	0.578	−0.049	0.256	0.442	0.003	0.085	0.487	−0.012	0.207	0.343
72	−0.009	0.216	0.548	−0.053	0.258	0.422	0.005	0.091	0.384	−0.012	0.210	0.311
96	−0.011	0.223	0.502	−0.051	0.265	0.378	0.005	0.090	0.394	−0.010	0.210	0.313
120	−0.029	0.242	0.368	−0.058	0.264	0.392	0.001	0.093	0.320	−0.015	0.206	0.361

The robust data set used in this study, the ML seeing catalogue, which contains selected variables from the GFS model output, is paired with the OT measurements provided by MD over the decade since their installation. The implemented model includes an ML module and data clustering module. In operational mode, the ML module learns the underlying associations between OT

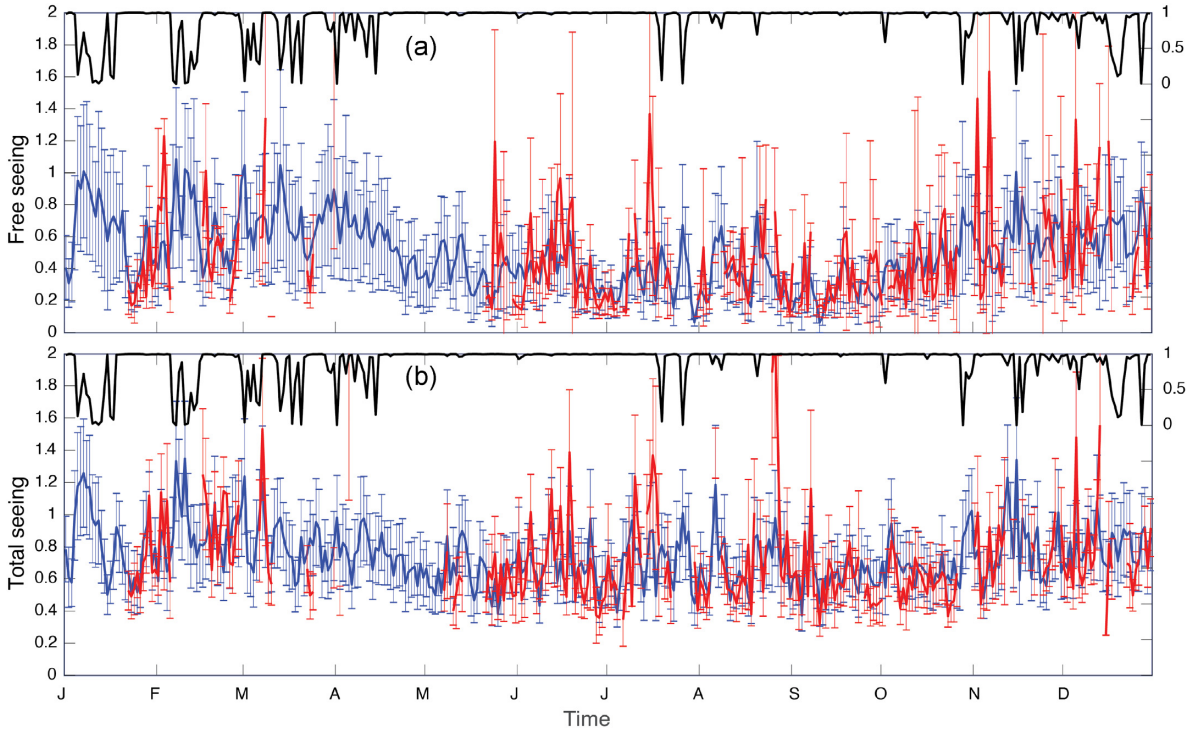

Figure 11. (a) 2020 timeseries of free atmospheric seeing as measured by MASS (red solid line) and as predicted by the ML algorithm (blue solid line); measured and predicted ± standard deviation is also shown. (b) 2020 timeseries of free atmospheric seeing as measured by DIMM (red solid line) and as predicted by the ML algorithm (blue solid line); measured and predicted ± standard deviation is also shown (thin error bars). Black lines show the probability for a given night to be a good or bad observing conditions night as predicted by the ML algorithm (*y*-axis on the right).

measurements and large-scale weather variables relevant for the summit of Maunakea from the ML seeing catalogue and provides a prediction for the average total and free atmospheric seeing, their standard deviations, and the expected status of each observing night (targets). The clustering module assigns the predictions for the nights in the ML seeing catalogue into four OT categories defined by the pairs: (L_G, L_{FREE}), (L_G, H_{FREE}), (H_G, L_{FREE}), (H_G, H_{FREE}).

The following is a summary of the conclusions from applying the ML model.

(i) The clustering module is able to group more than a decade of OT nightly data in four categories defined by pairing the following states: (i) weak (L_G) or (ii) strong (H_{FREE}) turbulence in the ground layer; and (iii) a calm (L_{FREE}) or iv) active (H_{FREE}) free atmosphere. The cluster's centroids defined by the *k*-means algorithm (Tables 3 and 4) identify average values of seeing that are largely consistent with the climatology of seeing for Maunakea (Lyman et al. 2020).

(ii) Analysis of the occurrences of the four categories by month shows that more turbulent nights and poorer seeing in the ground layer are likely during the winter season, with a smaller maximum in poor seeing during the middle of the summer when sporadic episodes of high surface winds and/or strong winds aloft are possible. The number of nights with better seeing in the free atmosphere peaks during summer as expected with the mean polar jet stream shifted northward away from Hawaii.

(iii) The occurrences of low OT events in both the ground layer and free atmosphere, analysed by month, are the largest of the four possible class pairings, peaking during the summer months. This result is consistent with the fact that Maunakea is a premier site for astronomy. Events during which both the ground and free atmospheric seeing are poor are relatively less common. These nights are more likely to occur during winter when the mean position of the polar jet stream is closest to Hawaii.

(iv) Using the most recent year of the *ML seeing catalogue* as a validation data set, the ML module of the implemented model shows the following.

(i) The wind speed on the model levels closest to the summit (650, 600, and 550 mb) are the most important factors in determining the total atmospheric seeing.

(ii) The potential temperature largely impacts OT prediction in the middle atmosphere, while wind speed at the jet level (250 mb) and close to the surface (650 mb) are important.

(iii) Moisture shows a minor to null influence, consistent with the observation that there is relatively little moisture above the summit of Maunakea, except during storms when MD observations are typically not available.

(iv) The ML algorithm shows good predictive skill, particularly in terms of the total atmospheric seeing, with a correlation coefficient of 0.67. The free atmospheric seeing prediction is slightly less accurate, with a correlation coefficient of 0.47.

(v) When the ML model is used in an operational setting and the GFS forecasts are used as input, the following results are found.

(i) The predicted behaviour of the total and free atmosphere OT and their standard deviations are well captured by the ML algorithm and their seasonality is reproduced in a way that is consistent with the observations.

(ii) Statistics show a good correlation coefficient (∼0.6) and good bias-corrected rms (∼0.2) for the predicted total seeing, and a fair ∼0.5 correlation coefficient and 0.25 bias-corrected rms for free atmospheric seeing for the forecast

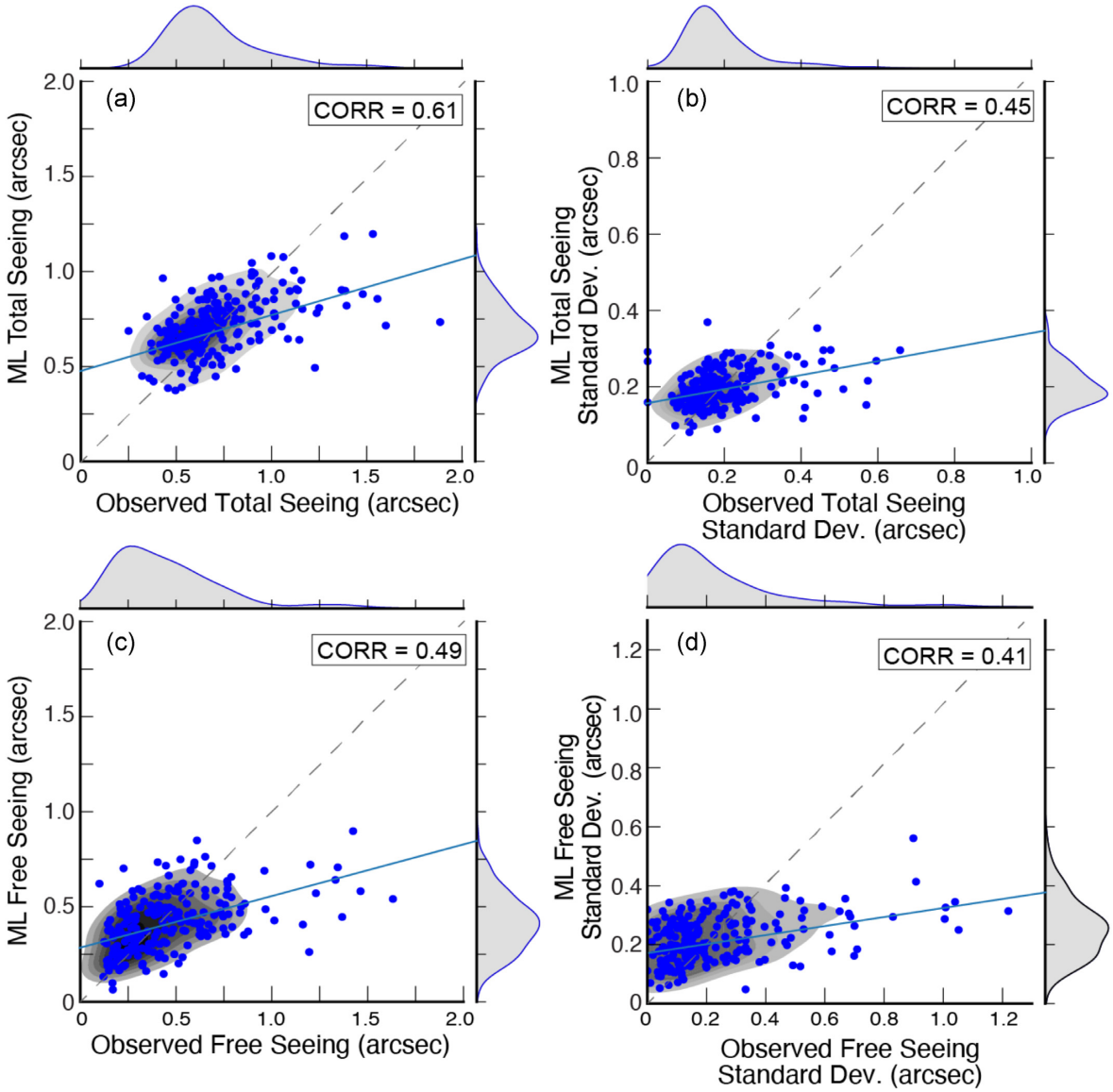

Figure 12. Scatterplots of the ML predicted versus observed: (a) total seeing and its (b) standard deviation; (c) free atmospheric seeing and its (d) standard deviation for 2020 in operational mode. Correlation coefficients are also shown.

hours 6, 12, and 18, which refer to the first observing night. Although a longer testing period is needed in order to draw the appropriate conclusions, the ML model seems to match the MKWC forecaster's performance (fig. 15 in Lyman et al. 2020) in predicting the nightly averaged OT behaviour at the summit of Maunakea.

5.1 Future work

Although the preliminary results presented in this study are very encouraging and the effectiveness of this approach are demonstrated by the presented statistics, there is room for improvement. For example, a multi-output regression model should be investigated as the targeted outputs are not necessarily independent of each other. Also, sensitivity studies to the ML algorithm's hyperparameters could be carried out and more sophisticated regressors could be tested. A further algorithm upgrade could envision more classifications ($k > 2$) classes in order to capture, for example, extremely low OT episodes. Finally, the constraints behind the *status* variable could be made less conservative to include more cases with higher relative humidity and the impact of these cases in the statistics studied. Moreover, it is recommended that a flag recording the status of the observing night be included in the instrumentation at the summit of Maunakea.

An ongoing parallel study is being carried out: four $TKE_{\min}$ profiles have been generated that correspond to the four main OT scenarios commonly observed at the Maunakea summit and identified by the clustering algorithm: (L_G, L_{FREE}); (L_G, H_{FREE}); (H_G, L_{FREE}); and (H_G, H_{FREE}). The goal of the new study is to dynamically initiate OT within the MKWC WRF model's OT parametrization scheme.

In summary, this paper describes the implementation of an ML model that has a two-fold purpose: (i) to mimic the process followed by the MKWC forecaster when issuing his/her daily OT forecast, and thus provide the MKWC forecaster with an additional predictive tool to anticipate the average total and free seeing and their projected variability and (ii) to allow for construction of a dynamic calibration

of the optical turbulence algorithm in WRF by selecting a TKE_{min} profile that corresponds to the appropriate OT category for a given weather pattern. The results for this latter study will be presented in a future paper.

ACKNOWLEDGEMENTS

Thanks go to the Maunakea Observatories for providing the MD database and to the CFHT staff for the operational support to the instruments and database, in particular to Marc Baril and Billy Mahoney. Thanks go to Nancy Hulbert for her graphics support and to May Izumi for her editing of the final manuscript. We would like also to acknowledge high-performance computing support from Cheyenne (doi:10.5065/D6RX99HX) provided by NCAR's Computational and Information Systems Laboratory, sponsored by the National Science Foundation. This research was supported by the National Science Foundation through the Maunakea Support Services.

DATA AVAILABILITY

The data underlying this article will be shared on reasonable request to the corresponding author.

REFERENCES

Businger S., McLaren R., Ogasawara R., Simons D., Wainscoat R. J., 2001, Bull. Am. Meteorol. Soc., 83, 858
Cherubini T., Businger S., Lyman R., Chun M., 2008a, J. Appl. Meteorol. Climatol., 47, 1040
Cherubini T., Businger S., Lyman R., 2008b, J. Appl. Meteorol. Climatol., 47, 3033
Cherubini T., Businger S., 2011, in Businger S., Cherubini T., eds, Seeing Clearly: The Impact of Atmospheric Turbulence on the Propagation of Extraterrestrial Radiation. VWB Publishing, Texas
Cherubini T., Businger S., 2013, J. Appl. Meteorol. Climatol., 52, 498
Chun M. et al. 2009, MNRAS, 394, 1121
Coulman C. E., 1985, ARA&A, 23, 19
Da Silva S. C., 2012, MS thesis, University of Hawaii at Mānoa
Kornilov V. et al. 2007, MNRAS, 382, 1268
Lyman R., Cherubini T., Businger S., 2020, MNRAS, 496, 4734
Masciadri E., Vernin J., Bougeault P., 1999, A&AS, 137, 185
Masciadri E., Jabouille J. P., 2001, A&A, 376, 727
Masciadri E., Lombardi G., Lascaux F., 2014, MNRAS, 438, 983
Milli J. et al., 2019. Nowcasting the turbulence at the Paranal Observatory, Preprint (arXiv:1910.13767v1)
National Centers for Environmental Prediction/National Weather Service/NOAA/U.S. Department of Commerce, 2015, NCEP GFS 0.25 Degree Global Forecast Grids Historical Archive. Research Data Archive at the National Center for Atmospheric Research, Computational and Information Systems Laboratory, Boulder, CO
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Skamarock W. C. et al., 2008, A Description of the Advanced Research WRF Version 3 (No. NCAR/TN-475+STR). University Corporation for Atmospheric Research, p. 113
Skidmore W. et al., 2009, PASP, 121, 1151
Tokovinin A., 2002, PASP, 114, 1156
Tokovinin A., Kornilov V., 2007, MNRAS, 381, 1179

This paper has been typeset from a TEX/LATEX file prepared by the author.