Statistical learning as a new approach for optical turbulence forecasting

Christophe Giordano, Alohotsy Rafalimanana, Aziz Ziad, Eric Aristidi, Julien Chabé, Yan Fanteï-Caujolle, Catherine Renaud

Statistical Learning as a new approach for optical turbulence forecasting

C. Giordano[a], A. Rafalimanana[a], A. Ziad[a], E. Aristidi[a], J. Chabé[b], Y. Fanteï-Caujolle[a], and C. Renaud[a]

[a]Laboratoire J.L. Lagrange, OCA, UCA, CNRS, UMR 7293, Parc Valrose 06108 Nice Cedex 2, France

[b]Université Côte d'Azur, OCA, CNRS, IRD, Géoazur, 2130 route de l'Observatoire, 06460 Caussols, France

ABSTRACT

For the new generation of Extremely Large Telescopes the knowledge of atmospheric turbulence conditions is become an information of primary importance to design and optimize all focal instrumentation. In the same way, the forecast of these atmospheric conditions is also of interest to allow both flexible scheduling and long term site testing. Until now we have used weather forecast tools coupled with turbulence models to predict turbulence conditions. In addition, we are developing a predictive statistical learning tool, using a large atmospheric database. Since 2015, the Calern Observatory hosts the Calern Atmospheric Turbulence Station (CATS) which measures during daytime and nighttime, ground meteorological conditions, vertical profiles of the C_n^2 and all relevant integrated parameters characterizing the optical turbulence. This large CATS database is used as input for our predictive statistical learning tool. This latter should take into account more closely the local specificities, seasonal variations and day/night transitions. A comparison between turbulence predictive models and statistical learning tools is presented and discussed.

Keywords: Turbulence - Atmospheric effects - Forecasting - Flexible Scheduling - Modeling - Statistical analysis - Machine learning

1. INTRODUCTION

It is now well established that astronomical observatories need meteorological and turbulence forecasting to plan observing night.[1] This is called the flexible scheduling. During the past years some studies raised[2–13] to bring an answer to the problem of the observation planning within the largest observatories, and for the next generation of Extremely Large Telescope. These studies used mainly a global circulation model such as the Mesoscale Non Hydrostatic model[14] (Meso-NH) or the Weather Research and Forecasting[15] (WRF) model to forecast the weather parameters such as the temperature, the pression, the relative humidity and the wind speed characteristics (modulus and direction). These studies used also different methods and/or models to deduce turbulence parameters from weather forecasting. Up to now, the mainly used method are either model derived form Tatarskii theory,[4,5] or empirical model coming from a statistical analysis of a large number of measurements.[9,13,16] Recently, authors of Ref. 1 showed the possibility to also use machine learning to bring a new solution for the short-terms prediction to improve the flexible scheduling.

In this paper, we focus on both the empirical method from Ref. 16 and on machine learning possibilities. Indeed, in Ref. 13 we have introduced a way to improve the empirical model derived from balloon radio-sounding (BDTM: Balloon Derived Turbulence Model) by introducing local statistics of a given site to take into account local specificities of this site. We call this upgrade the site learning (SL) method. This work has been done on the Calern Observatory (France) where a turbulence station (Calern Atmospheric Turbulence Station: CATS [17–20]) monitors, since 2015, the daytime and nighttime turbulence and meteorological conditions. The large database acquired during these years allows us to run statistical analysis and upgrade of the aforementioned empirical model. This database also is also used in this paper in a machine learning purpose such as in Ref. 1.

E-mail: cgiordano@oca.eu

Table 1. Domains configuration used for the WRF forecasting.

Domains	D01	D02	D03	D04
Central latitude [deg]	43.7546 North			
Central longitude [deg]	6.9206 East			
$\Delta x \times \Delta y$ [km]	27×27	9×9	3×3	1×1
Δh [km]	$0.015 \leq \Delta h \leq 1.246$			
Δt [minute]	10			
Topographic resolution $\Delta x = \Delta y$ [km]	0.927	0.927	0.927	0.09
Meteorological input data	Global forecast system (GFS) $\Delta x = \Delta y = 0.25deg$			

In this paper we will recall the SL method presented in Ref. 13, and we will show new results based on a largest sample. Then we will use machine learning algorithme above Calern Observatory to study the possibility to use short-term predictions in a complementarity with the mid-term predictions.

2. DATABASE PRESENTATION

2.1 On-site measurements

As presented in previous papers,[13,17–20] the Calern Observatory is dotted, since 2015 the CATS station which measures the whole set of meteorological and turbulence parameters thanks to a new generation of complementary instruments:

- the Generalized Differential Image Motion Monitor (GDIMM) instrument[18,21,22] measures the integrated parameters of the optical turbulence (Seeing, Isoplanatic Angle, Scintillation, Outer Scale, Coherence time) by observing bright stars.
- The Profiler of Moon Limb (PML) instrument[23–26] measures, during daytime and nighttime, the vertical profiles of the turbulence intensity C_n^2 and all related parameters (Seeing, Isoplanatic Angle, Scintillation, Outer Scale) by observing solar/lunar limbs.

CATS measurements authorizations are controlled by a weather station and an All-sky visible camera which respectively measure ground weather parameters (pressure, temperature, relative humidity wind speed modulus and direction, precipitation) and nighttime cloud cover. Therefore the CATS provide daytime and nighttime measurements in an autonomous way.

Both Site learning and machine learning tools will use either the whole CATS database or a smaller part of it.

2.2 Calern forecasting

Since the end of July 2019, we run daily forecasting for the next 48h with a temporal resolution of 10 minutes. These forecasts are visible in real time on the CATS website*. The configuration used for the WRF simulation is detailed in table 1. Up to the end of September 2020, we have forecasted conditions for around XXX days, which corresponds to more than XXX points. This set of data has been used for the site learning study presented in sections 3 and 5.1.

*https://cats.oca.eu/

3. SITE LEARNING PRESENTATION

As detailed in Ref. 13, the empirical model deduced by authors of Ref. 16 depends on statistics of measurements of the structure constant of the temperature fluctuations $C_T^2(h)$, the vertical gradient of the potential temperature $\chi(h)$ and the horizontal wind shear $s(h)$ following equations 1 to 4 hereafter:

$$C_T^2(h) = \phi(h).\chi(h).s(h)^{1/2} \tag{1}$$

$$\chi(h) = \frac{d\theta}{dh} \tag{2}$$

$$s(h) = \sqrt{\left(\frac{dU}{dh}\right)^2 + \left(\frac{dV}{dh}\right)^2} \tag{3}$$

$$\phi(h) = \frac{\langle C_T^2(h)\rangle_m}{\langle\chi(h)\rangle_m.\langle s(h)\rangle_m^{1/2}} \tag{4}$$

where h is the altitude and U and V are respectively the wind speed in East-West and South-North directions.

Following this equation, $\phi(h)$ is the statistical profile depending on available measurements. Ref. 13 used a sample of couple month of the CATS database corresponding to the same sample of WRF forecasting to test the effect of using local measurements as a constrain for the empirical model. This is called the Site Learning (SL) method. They have found a real improvement on turbulence predictions, and in this paper we extend this study to a largest sample covering almost a year of predictions, with and without upgrading the $\phi(h)$ profile.

The figure 1 shows the number of day per month where there are measurements from CATS station and predictions from WRF model. As it appears in this figure 1, we have measurements spread during the all seasons, except during the spring where there is a lack of observations. The SL method used essentially the measurements of the weather station and of the PML instrument. Therefore, considering this figure, our SL method will be more representative of the conditions of the summer and autumn periods.

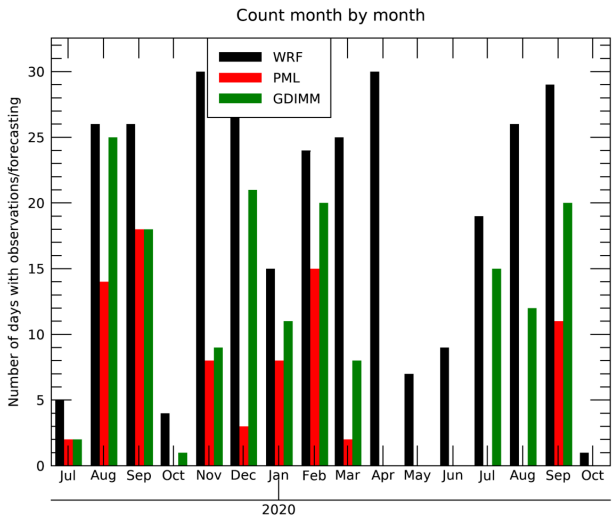

Figure 1. Distribution of the number of days per month where there are available data (measurements and/or forecasting), starting from July 2019. The black, red and green colors represent respectively WRF, PML and GDIMM data.

4. MACHINE LEARNING TOOL

4.1 Regression models used

The second part of our study concerns a shorter-term prediction using machine learning tools. Following the work of Ref. 1, we have decided to study the possibility of using machine learning for short-term predictions to bring a complementary information in our forecasting tool.

Our approach will consist of using Calern database as both training and test case with the following algorithms coming from the Scikit Learning library:[27]

- Ridge regression algorithm. It is similar to ordinary least squares regression but adding a penalty on the size of the coefficient to addresses their problems.
- Random forest regression[28] with 1000 decisions trees. It is based on the randomization of decision tree method.

4.2 Data preparation

To run machine learning algorithms, we have decided to use 2 hours of measurements to predict the seeing over the next 2 hours. We have divided our database between training sets and test sets following the scheme of figure 2.

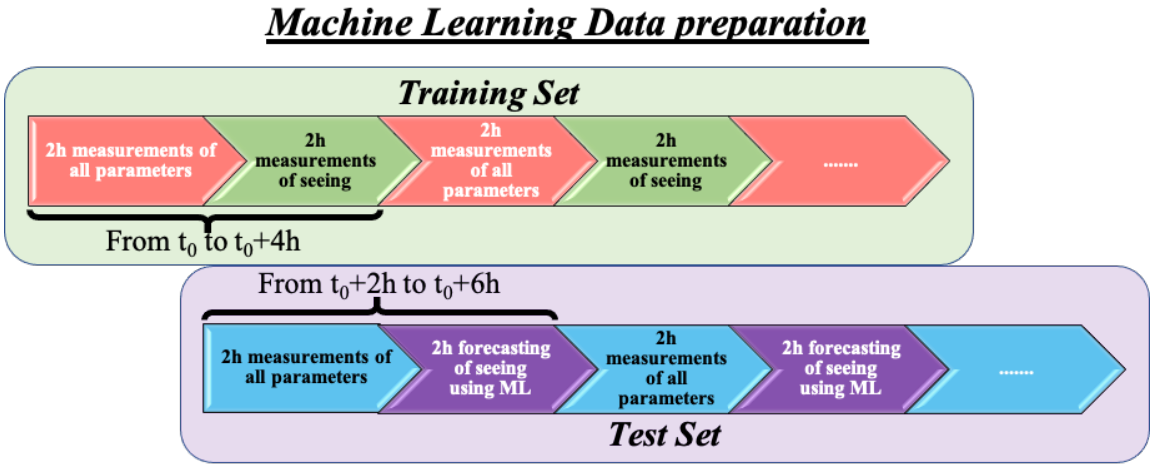

Figure 2. Scheme of the data preparation process for machine learning algorithms.

In our study the parameters used for the training sets and tests sets are composed of the following parameters:

- Day from the beginning of the year represented circularly as cosine and sine of day/365 to take into account periodicity of the year and seasonal effects.
- Hour represented circularly as cosine and sine of hour/24 to take into account periodicity of earth rotation and nighttime/daytime conditions.
- Ground meteorological parameters (wind speed modulus and direction, temperature, relative humidity, pressure) having a 1 minutes temporal resolution.
- Seeing having a temporal resolution of around 2 minutes.
- Isoplanatic angle having a temporal resolution of around 2 minutes.
- Coherence time having a temporal resolution of around 2 minutes.

Considering these parameters, we have decided to interpolate all data to a constant time step of 2 minutes. This work gives us a training set and test set composed respectively of 4297 and 4296 blocks of 2 hours. Each block is composed of a vector of 484 values:

- 1 value of cosine(day/365), sine(day/365), cosine(hour/24), sine(hour/24) per block
- 60 values per block of seeing, coherence time, isoplanatic angle, pressure, temperature, relative humidity, wind speed modulus and direction with a time step of 2 minutes.

During interpolation, missing values are considered as NaN. Then, before fitting with the two regression algorithms, we used the SK-learn Imputation transformer for completing missing values by mean values over the data set. The distribution of data used for this data preparation is shown in figure 3

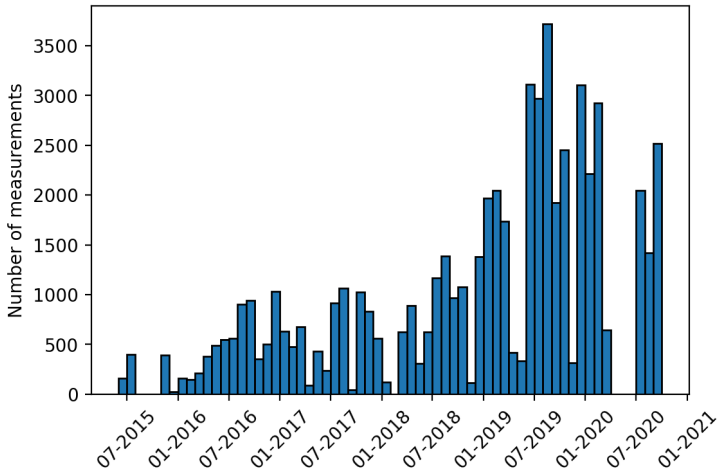

Figure 3. Distribution of data used during the machine learning data preparation from the year 2015.

5. RESULTS

5.1 Site learning

In this paper, we show the effect of site learning on the correlations between measurements and prediction but in a different way than in Ref. 13. Indeed, in figure 4, we plotted box showing the first quartile, the minima and the third quartile, and we added whiskers corresponding to minima and maxima. Each box has a width of 0.25 arcsec. This kind of plot gives interesting information about the correlations between measurements and prediction because we can see both the median values and dispersions of measurements within each bin.

In figure 4 we can easily see, the improvement brought by SL method on seeing predictions because the median values are closer to the y=x axis and the dispersions between first and third quartiles are smaller than with the initial BDTM method. In the PML cases (figure 4 bottom), improvements are visible for the small values of the seeing ($\varepsilon_0 < 1.5 arcsec$) where the median values get closer to the y=x axis and on all other values where the dispersion is reduced. Also, with SL method, WRF is now able to predict larger values of the seeing which can happen during daytime.

The seeing being only due to turbulence intensity (C_n^2), it is more impacted by lower layers which are more turbulent than higher. Moreover, this low layers conditions are the most difficult to predict because the complexity of the terrain and all related meteorological circulations above it are very hard to simulate with a mesoscale model (see Refs. 12, 13). Therefore, the differences and dispersions shown in this figure 4 can be explained by the difficulties to well retrieve low altitude C_n^2.

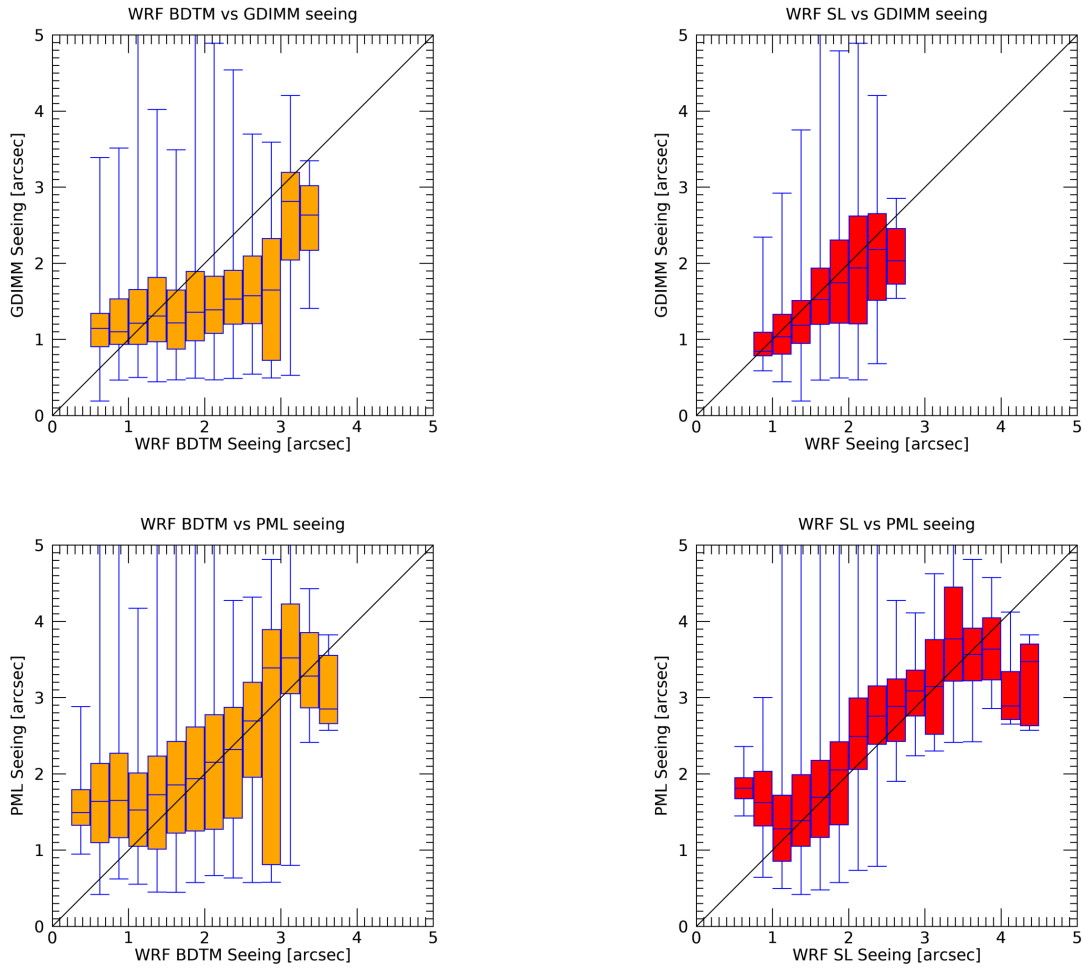

Figure 4. Scatter plot shown as box plot of the seeing forecasted using both BDTM (left) and SL (right) method and measured with GDIMM (top) and PML (bottom). The box corresponds to interval between first and third quartile. Within the box, the median values is plotted. The whiskers corresponds to extrema.

On figure 5, we show the same kind of plot than figure 4 but for the isoplanatic angle. Here also the improvement brought by SL method is important because the median is closer to the y=x axis for both GDIMM and PML comparisons. It is important to notice that the isoplanatic angle is a parameter mainly affected by high turbulent layers (see eq. 6 of Ref. 9). Indeed, the effect of altitude on the spatial coherence of the wavefront is larger when layers is higher, and even if the C_n^2 difference is low, the isoplanatic angle is impacted. Therefore, even if the forecasted C_n^2 at high levels is very well correlated to measurements, it remains a difference in the isoplanatic angle predictions.

5.2 Machine learning

As explained in section 4, we have used two different algorithms of machine learning in our study. Figure 6 shows the scatter plot (on the left) between seeing predictions found with machine learning and the real measurements from the GDIMM instrument. The right panel show the evolution of the Pearson correlation coefficient along the two hours of predictions. The first obvious results is that the Ridge regression seems to not work well enough in our case. However, regarding the Random forest regression, the correlation is much better and the Pearson correlation coefficient reaches values of around 0.8 at the beginning and decreases slowly during the time up to 0.2-0.25.

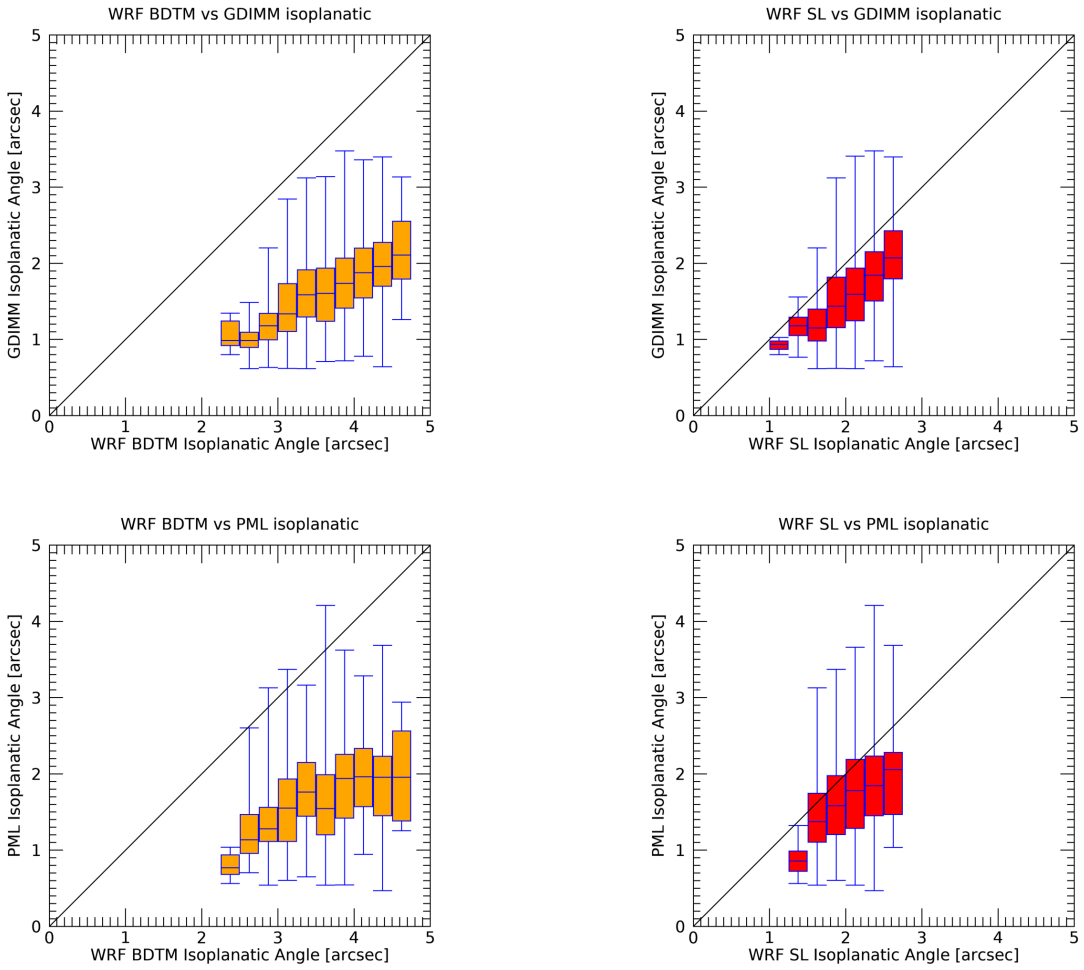

Figure 5. Same as figure 4 but for isoplanatic angle comparisons.

Usually, to deduce the seeing, we need either measurements on star and/or profiles of the C_n^2. Here, we only used ground parameters, and integrated coherence parameters. Therefore, these results are very encouraging for the beginning of our study.

6. CONCLUSION

In this paper we have presented the results of our Site Learning method computed on several months of predictions. As presented in Refs. 13 this SL method is a real improvement for our forecasting tool with respect to the initial BDTM method. Indeed, we have shown that the correlation on seeing prediction has been increased by 73% (PML) to 131% (GDIMM) and the absolute bias has been reduces by 23% (PML) to 30% (GDIMM). As detailed above, it remains dispersion within prediction which are mainly due to difficulties to well predict ground condition because of the terrain complexity. To overcome these issue, we will interest on the refinement of the resolution of the WRF grid. We will also interest to the possibility to ingest precise local conditions as input of the WRF forecasting to better take into account local specificities. After that, we will increase deeper the SL method to better adapt to the seasonal and diurnal conditions.

About our preliminary study on the machine learning tool, the first results are encouraging. We have, in a near future, to compare these predictions to WRF ones in order to see the impact of adding this kind of forecasting in our tool. In addition, we have to investigate other machine learning methods such as neural

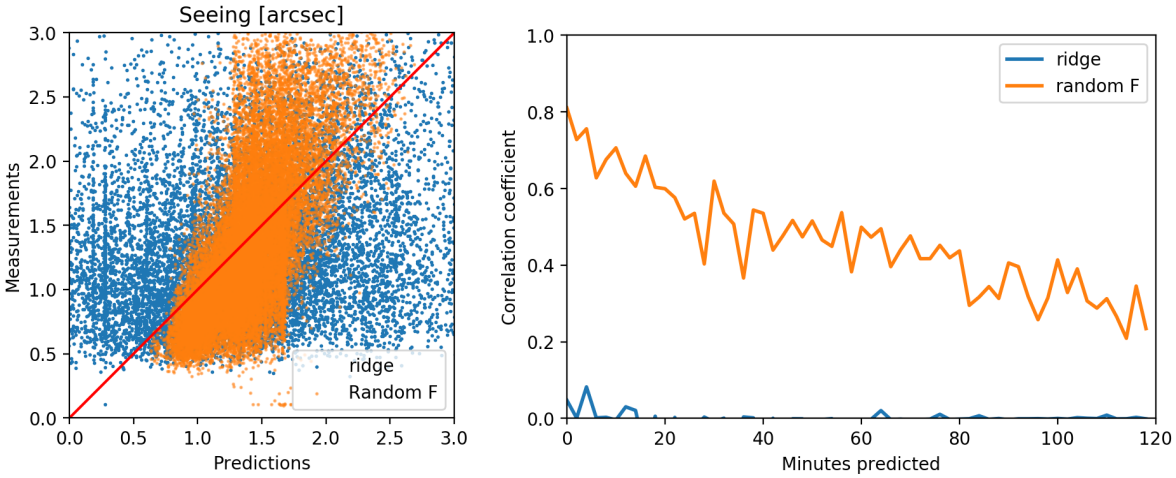

Figure 6. Scatter plot (left) and evolution of the Pearson correlation coefficient (right) during the 2 hours predicted for both algorithms (blue: ridge regression, orange: random forest regression) used in our study.

network to search for the better way to achieve our predictions. Finally, we need to introduce vertical profiles measurements of C_n^2 and other meteorological parameters during the data preparation. For this last part we will used data coming from PML monitor and from measurement campaign we are currently doing and which acquires data within the first hundreds of meters of the atmosphere.

REFERENCES

[1] Milli, J., Gonzalez, R., Fluxa, P. R., Chacon, A., Navarette, J., Sarazin, M., Pena, E., Carrasco-Davis, R., Solarz, A., Smoker, J., Martayan, C., Melo, C., Sedaghati, E., Mieske, S., Hainaut, O., and Tacconi-Garman, L., "Nowcasting the turbulence at the Paranal Observatory," *arXiv e-prints* , arXiv:1910.13767 (Oct. 2019).

[2] Coulman, C. E., Andre, J.-C., Lacarrere, P., and Gillingham, P. R., "The observation, calculation, and possible forecasting of astronomical seeing," *Publications of the Astronomical Society of the Pacific* **98**, 376–387 (Mar. 1986).

[3] Bougeault, P., de Hui, C., Fleury, B., and Laurent, J., "Investigation of seeing by means of an atmospheric mesoscale numerical simulation," *Applied Optics* **34**, 3481 (June 1995).

[4] Hach, Y., Jabiri, A., Ziad, A., Bounhir, A., Sabil, M., Abahamid, A., and Benkhaldoun, Z., "Meteorological profiles and optical turbulence in the free atmosphere with NCEP/NCAR data at Oukaïmeden - I. Meteorological parameters analysis and tropospheric wind regimes," *Monthly Notices of the Royal Astronomical Society* **420**, 637–650 (Feb 2012).

[5] Masciadri, E., Vernin, J., and Bougeault, P., "3D mapping of optical turbulence using an atmospheric numerical model. I. A useful tool for the ground-based astronomy," *Astronomy and Astrophysics Supplement* **137**, 185–202 (May 1999).

[6] Masciadri, E., Lascaux, F., Turchi, A., and Fini, L., "Optical turbulence forecast: ready for an operational application," *Monthly Notices of the Royal Astronomical Society* **466**, 520–539 (Apr 2017).

[7] Cherubini, T., Businger, S., Lyman, R., and Chun, M., "Modeling Optical Turbulence and Seeing over Mauna Kea*," *Journal of Applied Meteorology and Climatology* **47**, 1140 (2008).

[8] Cherubini, T., Businger, S., and Lyman, R., [*An operational perspective for modeling optical turbulence. Seeing Clearly: The Impact of Atmospheric Turbulence on the Propagation of Extraterrestrial Radiation.*], 165–182, S. Businger and T. Cherubini, Eds., VBW Publishing (2011).

[9] Giordano, C., Vernin, J., Vázquez Ramió, H., Muñoz-Tuñón, C., Varela, A. M., and Trinquet, H., "Atmospheric and seeing forecast: WRF model validation with in situ measurements at ORM," *Monthly Notices of the Royal Astronomical Society* **430**, 3102–3111 (Apr. 2013).

[10] Giordano, C., Vernin, J., Trinquet, H., and Muñoz-Tuñón, C., "Weather Research and Forecasting prevision model as a tool to search for the best sites for astronomy: application to La Palma, Canary Islands," *Monthly Notices of the Royal Astronomical Society* **440**, 1964–1970 (May 2014).

[11] Giordano, C., *Prédiction et optimisation des techniques pour lâĂŹobservation aÌĂ haute reÌĄsolution angulaire et pour la future génération de très grands télescopes*, PhD thesis, Université de Nice Sophia Antipolis, Nice, France (2014).

[12] Giordano, C., Rafalimanana, A., Ziad, A., Aristidi, E., Chabé, J., Fantei-Caujolle, Y., and Renaud, C., "Optical turbulence forecasting and comparisons with daytime and nighttime measurements," in [*Proceedings of the sixth AO4ELT Conference*], (2019).

[13] Giordano, C., Rafalimanana, A., Ziad, A., Aristidi, E.and Chabé, J., Fantéi-Caujolle, Y., Renaud, C., and Giordano, C., "Contribution of statistical site learning to improve optical turbulence forecasting," *accepted with moderate revision Monthly Notices of the Royal Astronomical Society* (2020).

[14] Lafore, J. P., Stein, J., Asencio, N., Bougeault, P., Ducrocq, V., Duron, J., Fischer, C., Héreil, P., Mascart, P., Masson, V., Pinty, J. P., Redelsperger, J. L., Richard, E., and Vilà-Guerau de Arellano, J., "The Meso-NH Atmospheric Simulation System. Part I: adiabatic formulation and control simulations," *Annales Geophysicae* **16**, 90–109 (Jan. 1998).

[15] Skamarock, W. C., Klemp, J. B., Dudhia, J., Gill, D., Liu, Z., Berner, J., Wang, W., Powers, J., Duda, M., Barker, D., and Huang, X.-Y., "A Description of the Advanced Research WRF Version 4," tech. rep., NCAR (2019).

[16] Trinquet, H. and Vernin, J., "A statistical model to forecast the profile of the index structure constant C_N^2," *Environmental Fluid Mechanics* **7**, 397 (Sept. 2007).

[17] Chabé, J., Ziad, A., Fantéï-Caujolle, Y., Aristidi, É., Renaud, C., Blary, F., and Marjani, M., [*The Calern atmospheric turbulence station*], vol. 9906 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 99064Z (2016).

[18] Aristidi, E., Fantei-Caujolle, Y., Chabé, J., Renaud, C., Ziad, A., and Ben Rahhal, M., "Turbulence monitoring at the Plateau de Calern with the GDIMM instrument," in [*Proceedings of the SPIE*], *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* **10703**, 107036U (July 2018).

[19] Ziad, A., Chabé, J., Fantei-Caujolle, Y., Aristidi, E., Renaud, C., and Ben Rahhal, M., "CATS: an autonomous station for atmospheric turbulence characterization," in [*Proceedings of the SPIE*], *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series* **10703**, 107036L (July 2018).

[20] Ziad, A., Chabé, J., Fantei-Caujolle, Y., Aristidi, E., Renaud, C., and Giordano, C., "The Calern Atmospheric Turbulence Station," in [*Proceedings of the sixth AO4ELT Conference*], (2019).

[21] Aristidi, E., Fanteï-Caujolle, Y., Ziad, A., Dimur, C., Chabé, J., and Roland, B., [*A new generalized differential image motion monitor*], vol. 9145 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 91453G (2014).

[22] Aristidi, E., Ziad, A., Chabé, J., Fantéi-Caujolle, Y., Renaud, C., and Giordano, C., "A generalized differential image motion monitor," *Monthly Notices of the Royal Astronomical Society* **486**, 915–925 (Jun 2019).

[23] Ziad, A., Blary, F., Borgnino, J., Fanteï-Caujolle, Y., Aristidi, E., Martin, F., Lantéri, H., Douet, R., Bondoux, E., and Mekarnia, D., "PML/PBL: A new generalized monitor of atmospheric turbulence profiles," in [*Proceedings of the Third AO4ELT Conference*], Esposito, S. and Fini, L., eds., 126 (Dec. 2013).

[24] Blary, F., Ziad, A., Borgnino, J., Fantéï-Caujolle, Y., Aristidi, E., and Lantéri, H., [*Monitoring atmospheric turbulence profiles with high vertical resolution using PML/PBL instrument*], vol. 9145 of *Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series*, 91453H (2014).

[25] Aristidi, E., Ziad, A., Fantéï-Caujolle, Y., Chabé, J., Giordano, C., Renaud, C., and Lantéri, H., "Monitoring daytime and nighttime optical turbulence profiles with the PML instrument," in [*Proceedings of the sixth AO4ELT Conference*], (2019).

[26] Chabé, J., Aristidi, E., Ziad, A., Lantéri, H., Fanteï-Caujolle, Y., Giordano, C., Borgnino, J., Marjani, M., and Renaud, C., "PML: a generalized monitor of atmospheric turbulence profile with high vertical resolution," *Applied Optics* **59**, 7574 (Sept. 2020).

[27] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E., "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research* **12**, 2825–2830 (2011).

[28] Breiman, L., "Random forests," *Machine Learning* **45**(1), 5–32 (2001).