Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY

MNRAS **523,** 3487–3502 (2023) https://doi.org/10.1093/mnras/stad1552

Optical turbulence forecasts at short time-scales using an autoregressive method at the Very Large Telescope

E. Masciadri ,★ A. Turchi and L. Fini

INAF Osservatorio Astrofisico di Arcetri, Largo Enrico Fermi 5, I-50125 Florence, Italy

Accepted 2023 May 18. Received 2023 May 5; in original form 2022 June 15

ABSTRACT

In this study we demonstrate that we can provide forecasts of all the main astroclimatic parameters (seeing, wavefront coherence time, isoplanatic angle, and ground-layer fraction) on time-scales of 1 and 2 h (the most critical ones for the service mode) with a root-mean-square error (RMSE) that is smaller than or, at worst, comparable to the instrumental uncertainty (i.e. the standard deviation between instrument estimates). The seeing RMSE is 0.08 arcsec. Results are achieved thank to the use of the autoregressive method (AR) in our automatic forecast system and the study is applied to the Very Large Telescope (VLT). The AR method is a hybrid method taking into account forecasts of a non-hydrostatical mesoscale model jointly with real-time observations made in situ. We demonstrate that the AR method allows an improvement in forecast performance of roughly a factor of three or more with respect to the standard forecasts at a long time-scale (beginning of the afternoon for the coming night), depending on the parameter and the time-scale (1 and 2 h). The AR method also allows roughly a factor of two gain with respect to prediction by persistence. We also show that the AR method provides significantly better performance than a random-forest machine-learning algorithm. An extended analysis of the AR performance is provided following different strategies. Results achieved in this study are therefore very promising and tell us that we can provide real assistance to the service mode of the VLT instrumentation supported by adaptive optics systems.

Key words: turbulence – atmospheric effects – instrumentation: high angular resolution – method: data analysis – methods: numerical – site testing.

1 INTRODUCTION

Over the last few decades, new methods of modelling and forecasting optical turbulence (OT) have been very important for ground-based astronomical observations using adaptive optics or interferometry. We rely on numerical models, together with measurements, to characterize the turbulence conditions affecting the adaptive optics performance. On one side, real-time measurements offer the advantage of giving a direct estimate of OT parameters; however, they provide only local information relative to a single point (in the case of radiosoundings or mast) or a single line of sight (in the case of optical instruments). Atmospheric numerical models offer instead the advantage of modelling the whole volume around the telescope and they can provide 3D C_N^2 maps (either in the present or at a past time) at whatever location on Earth is required, without the necessity of putting expensive instruments in place for a site-testing campaign.

Models are certainly less expensive than instrumentation and they can be more easily used for OT characterization over long periods. However, as reported in Masciadri, Turchi & Martelloni (2019), all models require a calibration. If we relax the constraints on the calibration, the model performance degrades too.[1] The right perspective is therefore not to focus on identifying which of measurements and modelling is the better method, but to use measurements and modelling synergically.

The real key characteristic of modelling (not accessible to measurements) is the property of models to provide forecasts that allow us to project our knowledge into the future. This property is crucial for the science operation of current top-class telescopes, as well as next-generation telescopes, i.e. Extremely Large Telescopes (ELTs). This is because telescope optimization (for the management of both scientific programmes and instrumentation) can be greatly enhanced by advance knowledge of the atmospheric conditions. Modelling is a more suitable tool for this goal than measurements, for the reason previously explained. A large part of top-class telescope instrumentation is supported by adaptive optics (AO), and a few ELTs have even been conceived to be adaptive telescopes (such as the Giant Magellan Telescope (GMT) and the ELT[2]). This tell us that an OT forecast is crucial for ground-based astronomy supported by AO. All AO systems and their performance are indeed strongly dependent on turbulence constraints, which become more and more restrictive if we go to short wavelengths. Also, different features of the turbulence (C_N^2, seeing ε, isoplanatic angle θ_0, and wavefront coherence time τ_0) impact at different levels on the

★ E-mail: elena.masciadri@inaf.it

[1] Therefore a universal calibration will in general be less performant than a calibration dedicated to a site.

[2] Unluckily, the name of the ELT that will be operated by ESO is the same as that of the general term indicating the class of telescopes having a diameter of 30-40 m.

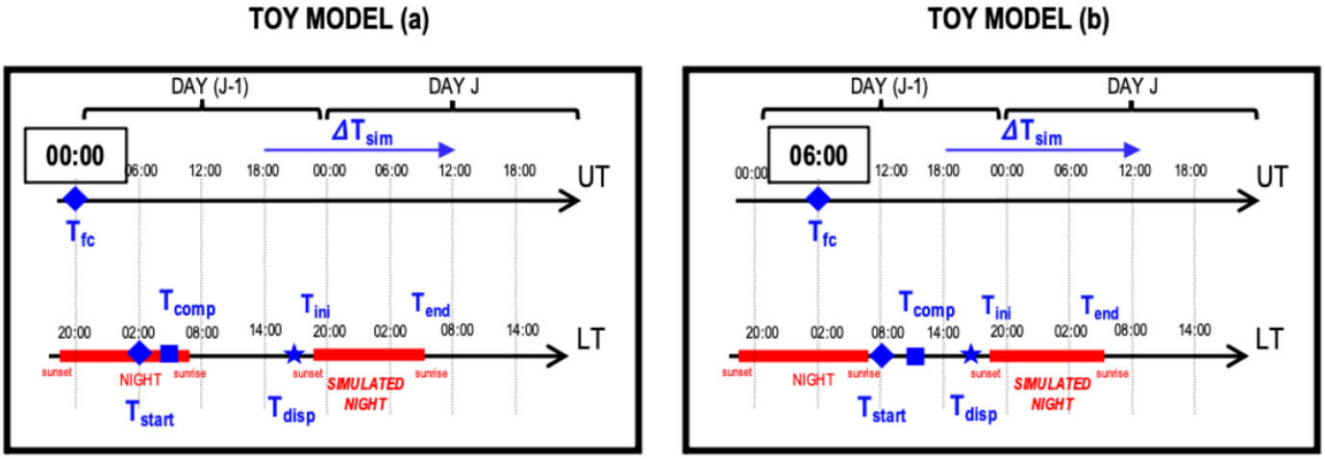

Figure 1. Toy models showing one of the possible forecast configurations. Left: initialization forecast data calculated at 00:00 UT of the day ($J-1$); right: initialization forecast data calculated at 06:00 UT of the day ($J-1$). The example shown in this figure reports two time axes. On top we have Universal Time (UT); on the bottom we have Local Time (LT). Depending on the location of the site, we can imagine replacing the time axis on the bottom with the local time typical of South America, Africa, Europe, etc.

different typologies of AO systems (Single Conjugated Adaptive Optics (SCAO), Laser Tomography Adaptive Optics (LTAO), Multi-Conjugated Adaptive Optics (MCAO), or Multi-Object Adaptive Optics (MOAO)). The isoplanatic angle, for example, is of little interest for a SCAO system but becomes critical for a wide-field adaptive optics system such as a Ground Layer Adaptive Optics (GLAO) (Rabien et al. 2019) or MCAO system (Neichel et al. 2014; Rigaut et al. 2014). The parameter τ_0 is critical for those instruments dedicated to high-contrast imaging (HCI) supported by SCAO systems (Spectro-Polarimetric High-contrast Exoplanet REsearch instrument (SPHERE), Gemini Planet Imager (GPI): Cantalloube et al. 2020; Madurowicz et al. 2018), those instruments dedicated to HCI running at short wavelengths such as the System for coronography at High order Adaptive optics from R to K band (SHARK-VIS) (Pedichini et al. 2016), also supported by a SCAO system, or those instruments dedicated to wide-field observations supported by a MCAO system, such as the MCAO-Assisted Visible Imager and Spectrograph (MAVIS: Rigaut et al. 2020). Scintillation can be considered negligible with respect to perturbations of the phase in planet finders with direct imaging (Masciadri, Feldt & Hippler 2004a), but it might become more and more important for higher levels of AO correction. Scintillation equally plays a role in the asymmetry of the 'butterfly effect' that strongly reduces the contrast (Cantalloube et al. 2018) in high contrast imaging (HCI). An extensive description of the challenging goals related to the forecast of OT and all derived astroclimatic parameters in ground-based astronomy has already been provided (Masciadri, Lascaux & Fini 2013; Masciadri, Martelloni & Turchi 2020). We refer the readers to the aforementioned articles for a complete analysis of this topic. Besides this, we also note that the OT forecast is also useful to decrease the time spent in out-of-constraints observations due to turbulence, as it has been quantified at the VLT and enables more aggressive short-term scheduling with well-estimated risks (Milli et al. 2019).

So far the typical questions we ask in the astronomical context are whether we are able to forecast OT, which atmospheric model is preferable to use in certain conditions, or what performance can be achieved by a certain method. At present we can affirm that the forecast of OT, i.e. C_N^2 and integrated astroclimatic parameters, is a reality. The first exhortation to use atmospheric models to reconstruct OT in the atmosphere was given by Coulman et al. (1986). The first attempts to model and forecast C_N^2 with pure atmospheric models were carried out in the late 1990s (Masciadri, Vernin & Bougeault 1999a,b). Several others followed in the next decades. Two main methods are employed for OT forecasts obtained with pure atmospheric models, as follows.

(a) The numerical approach, in which C_N^2 is parametrized using the prognostic turbulent kinetic energy equation. Studies related to this approach include Masciadri et al. (1999a, 2013); Masciadri & Egner (2006); Masciadri, Vernin & Bougeault (2001); Masciadri, Avila & Sanchez (2002, 2004b); Cherubini et al. (2008), Cherubini & Businger (2013); Lascaux, Masciadri & Hagelin (2009, 2010, 2011); Hagelin, Masciadri & Lascaux (2011); Masciadri, Lascaux & Fini (2017); Basu et al. (2020); Lyman, Cherubini & Businger (2020).

(b) The analytical approach, in which C_N^2 is expressed as a function of temperature and wind speed using algorithms mainly obtained with empirical fits or physical considerations. Studies related to this approach are, for example, Dewan (1993); VanZandt et al. (1978); Trinquet & Vernin (2007); Ruggiero & De Benedictis (2002); Ye (2011); Giordano et al. (2013); Osborn & Sarazin (2018); Wu (2020).

Both approaches (a) and (b) imply a calibration (even if the calibration can be different depending on the approach used). Masciadri & Jabouille, (2001) was the first calibration method proposed in the literature, but we do not enter into the details of the model calibration as this is not a topic of this work. More recently, studies implying the use of filtering techniques and the machine-learning approach have started to appear (Turchi, Martelloni & Masciadri 2018; Milli et al. 2019; Masciadri et al. 2020; Bolbasova, Andrakhanov & Shikhovtsev 2021; Cherubini, Lyman & Businger 2022), even if with a very wide spectrum of applications.

So far, not too much attention has been dedicated to the definition of the forecast time-scale (FTS), at least in the astronomical context, and this represents a great limitation, because if we do not consider the FTS we cannot really appreciate the goodness of a forecast, nor compare different approaches. If we do not define the configuration of a forecast, it is meaningless to quantify its efficiency. But what do we mean by FTS?

Many different typologies of *time* and *time intervals* play a role when a forecast of a parameter related to the Earth's atmosphere is performed. Toy models in Fig. 1 show the most important ones.

The description reported here is suitable for applications to astronomy. Looking at Fig. 1(a) we have the following.

(i) The time at which the initialization data coming from a General Circulation Model (GCM) are calculated: $T_{\rm fc}$.

(ii) The time at which the simulation that is supposed to deliver the OT forecast starts: $T_{\rm start}$. This is a 'technical time' required to receive initialization data from the GCMs to be used to perform forecasts with a mesoscale model. In the case of the GCM of the European Centre for Medium-Range Weather Forecasts (ECMWF), $T_{\rm start}$ is 6 hours.

(iii) The time at which the outputs of the forecasts are displayed, therefore the time at which we are aware of the information. This corresponds to the time at which the observatory requires forecasts to be delivered: $T_{\rm disp}$.

(iv) The time at which the computation of the forecast ends, which we call $T_{\rm comp}$, which has to respect the condition $(T_{\rm comp} - T_{\rm start}) \leq (T_{\rm disp} - T_{\rm start})$. See Appendix A for further information.

(v) The simulated time, which is a time interval that is forecasted in the future: $\Delta T_{\rm sim}$.

(vi) The time of post-processing of forecast data $T_{\rm pp}$, which in our case is basically negligible, as we are talking about a few seconds or minutes.

(vii) The temporal sampling of the forecast: $T_{\rm freq}$. This is the temporal sampling with which the forecast is delivered and displayed. In our case it is of the order of a few seconds or minutes, depending on the parameter.

(viii) The *forecast time-scale* (FTS). We define the FTS as the interval between the time at which the initialization forecast is calculated ($T_{\rm fc}$) and the time the information refers to. This means that the FTS is included between $(T_{\rm ini} - T_{\rm fc})$ and $(T_{\rm end} - T_{\rm fc})$.[3] If we consider that the night is included in the range $[T_{\rm ini}, T_{\rm end}]$ and we assume for example that $T_{\rm ini}$ = 19:00 local time (LT) on day $(J-1)$ and $T_{\rm end}$ = 05:00 LT on day J, then the FTS is in the range [23–33 h] in the case of Fig. 1(a).
In our analysis, we consider the FTS related to the whole night (i.e. [23–33 h]) as a unique FTS. We assume, in other words, that the accuracy of the forecasts all through the night in the [23–33 h] range is the same – see Section 4. Toy model (b) shows an example in which the FTS = [17–27 h] as $T_{\rm fc}$ is shifted 6 h ahead. As a first approximation, the shorter $(T_{\rm ini} - T_{\rm fc})$, the better we are (see Appendix A for further details). We use 'standard configuration' to denote the forecast obtained using a configuration such as that shown in Fig. 1 and delivered at $T_{\rm disp}$ with respect to the whole coming night.

(ix) We can also define the forecast time-scale related to the displayed time ($\rm FTS_{disp}$) as the difference between the time at which we are aware of the information and the time the forecast refers to, i.e. $(T_{\rm ini} - T_{\rm disp})$ and $(T_{\rm end} - T_{\rm disp})$. This parameter is not related to the forecast performance, but tells us how much in advance we know the atmospheric conditions. $\rm FTS_{disp}$ is relevant because it determines constraints in the model configuration. In the case of Fig. 1(a), $\rm FTS_{disp}$ is in the [2–12 h] range.

The configurations shown in Fig. 1 are certainly not unique possible forecast configurations. The selection of the operational forecast configuration depends on many factors. It is usually the result of a trade-off between the cost of hardware architecture and the output model performance, as well as the model configuration and the final goal of the application. It is enough here to know that the values of the different $T_{\rm X}$ and FTS can change in different contexts and in different forecast systems. For this reason it is important to describe the forecast scheme used in a study or in an operational system so that readers can appreciate the results/outputs. We refer the readers to Appendix A to avoid misunderstandings. We finally highlight that this is not supposed to be an operational meteorological forecast system, such as regional meteorological forecast systems, but an automatic operational forecast system for astronomical purposes.

Masciadri et al. (2020) recently proposed a method based on the principle of autoregression (AR) that is able to improve the forecast performance strongly on short time-scales (1 or 2 h). This improvement is relevant for the seeing and all key atmospheric parameters (temperature, relative humidity, wind speed, and direction). The gain is a factor in the [2.7, 5] range. For the precipitable water vapour (PWV), it has been proven that the gain can achieve the order of ~ 8 (Turchi et al. 2020). We will hereafter call this method simply AR. Time-scales of 1 or 2 h are certainly the most critical ones for science operation in the ground-based astronomy supported by AO.

Among the most relevant results obtained by Masciadri et al. (2020), it has been demonstrated that, for the seeing forecasts at 1 h, the root-mean-square error (RMSE) = 0.1 arcsec. Also, the probability of detecting the seeing below the first tertile of a climatological distribution ($\rm POD_1$ – see Appendix B) is 99 per cent on the same time-scale of 1 h. In the same work, it has also been proven that the AR method provides better performance than prediction by persistence for the seeing and the atmospheric parameters. The Masciadri et al. (2020) study was performed on the Mt Graham (Arizona), site of the Large Binocular Telescope (LBT) and the AR method has been implemented in the Advanced LBT Turbulence and Atmosphere (ALTA) Center system,[4] which is, to our knowledge, the first center delivering automated forecasts at long and short time-scales in the astronomical context. The AR method is based on simultaneous use of numerical forecasts from the Astro-Meso-NH model and real-time measurements (see Masciadri et al. 2020 for details). The role of in situ observations is basically to correct the forecast trend so as to improve the forecast performance. The AR method can therefore be applied only if real-time measurements are available. LBT is equipped with a Differential Image Motion Monitor (DIMM) for dedicated seeing measurements, but so far no optical turbulence vertical profilers are accessible in situ, even if the implementation of a vertical profiler is planned in the near future. This prevented us from applying the AR method to astroclimatic parameters different from seeing at LBT.

The goal of this work is to demonstrate that, for a specific forecast configuration, the AR method can provide, at the present time, forecasts with high accuracy on a time-scale of 1 and 2 h not only for the seeing (as done by Masciadri et al. 2020) but for *all* the most relevant astroclimatic parameters: seeing, wavefront coherence time (τ_0), isoplanatic angle (θ_0), and ground-layer fraction (GLF). This study is applied to Cerro Paranal, site of the Very Large Telescope (VLT), with the goal of providing forecasts for the queue scheduling of the VLT Service Mode. We will therefore use model outputs obtained with a configuration conceived for real-time forecasts. To achieve our goal we used, besides an atmospheric model, real-time measurements from the European Southern Observatory (ESO) Ambient Condition monitors. This allowed us to perform a complete analysis of the AR approach using statistical operators such as BIAS,

[3] $T_{\rm ini}$ and $T_{\rm end}$ are the times at which the night starts and ends.

[4] http://alta.arcetri.inaf.it

RMSE, SD, and the probability of detection POD_i retrieved from the contingency tables (see Appendix B).

We intend to show that, for *all* astroclimatic parameters, the forecast performance achievable at the present time with our system on these time-scales is already better than or comparable with the accuracy achievable with measurements and better than the prediction by persistence. This is a strong argument to guarantee efficient assistance in the science operation of instrumentation supported by AO at the VLT.

A further promising element is the fact that results presented in this work do not represent our best available results. A couple of elements are not considered in this work, as they will require important modifications of the data, but they might provide improvements in our system. (1) We recently evaluated the possibility of considering a more favourable forecast configuration for the operational version at the VLT that allows us to gain 6 h. This means shifting towards the right T_{fc} of 6 h. See Appendix A for further elements. (2) A new model calibration might be implemented with a new release of measurements (Release 2019B: Butterley et al. 2020) obtained with Stereo-SCIDAR. This might lead to an improvement of results that we have not yet estimated. We highlight these two elements just to illustrate that the goal of this work is not to show our best results but to demonstrate/show that, with this beta version, it is already possible to provide forecasts with an accuracy that is already sufficiently good to be useful for the science operation of top-class telescopes and, in particular, the VLT.

In Section 2 we provide a description of real-time measurements used as a reference to quantify the forecast performance. In Section 3 we describe the model and the forecast configuration. In Section 4 we present results on model performance obtained with the support of statistical operators (BIAS, RMSE, SD, contingency tables, POD_i). In Section 5 we include in our analysis the performance obtained with a random-forest machine-learning algorithm. In Section 6 we offer conclusions and discuss perspectives.

2 OBSERVATIONS

For this study we used archival VLT observations made in order to monitor optical turbulence. ESO observations have been stored in a public repository[5] since 2016 April. Observations we use refer to a DIMM and a Multi Aperture Scintillation Sensor (MASS). DIMM provides measurements of the integrated content of turbulence along the line of sight, which are corrected by the zenith angle. DIMM estimates are based on the measurement of the variance of the fluctuations of angles of arrival of the wavefront coming from a single star and passing through two small holes separated by a distance d within the pupil of a small-size telescope (Sarazin & Roddier 1990). The telescope pupil size is typically around 30 cm, the hole size $D \sim 6$ cm, and the distance $d \sim 14$ cm. Values of D, d, and the pupil size of the small telescope can change depending on the prototype provided $d \geq 2 \times D$. ESO-DIMM is a 11 arcsec Celestron on an Astelco NTM500 direct-drive mount. The seeing ε is retrieved from the longitudinal σ_l^2 and transversal σ_t^2 variances of the fluctuations of the angles of arrival. The temporal sampling of measurements is around 1 min.

MASS is a vertical profiler and has been introduced in the astronomical context by Kornilov et al. (2003) with a slightly modified version of a technique proposed by Ochs et al. (1976). It reconstructs the turbulence stratification in six layers distributed within 20 km above the ground, starting from the scintillation indices (four normal and six differential indices) of scintillation maps produced by single stars on a set of four small concentric and circular apertures selected on the telescope's pupil. Scintillation is measured by photomultipliers. The vertical stratification is obtained by fitting a set of measured scintillation indices with a model having a small and fixed number of turbulent layers. The layers are located at six heights: 0.5, 1, 2, 4, 8, and 16 km above the ground. The scintillation indices depend on the integral of C_N^2, the turbulence spectrum, the Fresnel diffraction term, and the pupil filter. The product of the turbulence spectrum times the Fresnel diffraction term times the pupil filter constitutes the weighting function (WF) $W(h)$. Besides stratification of the optical turbulence (C_N^2 or J profiles) at low vertical resolution, MASS also provides values of the isoplanatic angle (θ_0) and wavefront coherence time (τ_0). Different versions of DIMM and MASS have been used at the VLT at different epochs. In this work we refer to the most recent versions of MASS and DIMM that have been in operation since 2016 April. The temporal sampling of measurements is not constant, but we can say it is mostly of the order of 1 or 2 minutes. The two instruments are physically joined together. The MASS–DIMM instrument is based on a new version of the MASS software called ATMOS[6]that, in the reconstruction process of the vertical profile, also includes measurements coming from the DIMM, allowing more precise estimates (Kornilov et al. 2014). It operates in robotic mode on top of a 7-m high tower (Fig. 2) located to the north of UT4. For further information, we refer the reader to the ESO documentation.[7]

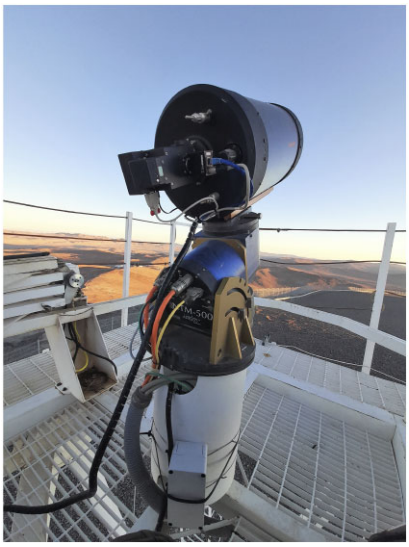

Figure 2. ESO MASS-DIMM located at Cerro Paranal, site of the VLT in Chile. The instrument is placed on top of a 7-m high tower. Credit: Angel Otarola, ESO.

In this work we also treated observations from Stereo-SCIDAR (SS: Butterley et al. 2020) to quantify the measurement accuracy. This instrument requires to be installed on a telescope of a diameter of at least 1 m and it is therefore not suitable for use as a monitor. However, ESO has so far collected measurements with Stereo-SCIDAR on a reasonably rich statistical sample of nights which is useful for our analysis. Stereo-SCIDAR (Sheperd et al. 2014) is a

[5]http://archive.eso.org/cms/eso-data/ambient-conditions.html

[6]ATMOS 2.9 version and above.

[7]Document Number: ESO-281474 – Release on 2017 November 7.

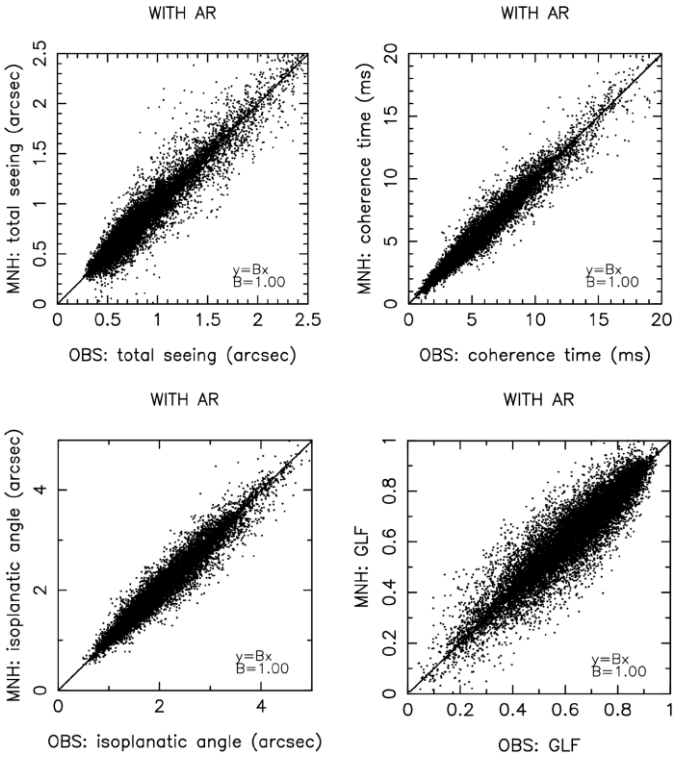

Figure 3. Scatter plot of observed and forecast parameters using the AR method on a time-scale of 1 h: seeing (ε) (top left), wavefront coherence time (τ_0) (top right), isoplanatic angle (θ_0) (bottom left), and GLF × 100 (per cent) (bottom right). B is the slope of the regression line. For the seeing, observations are taken from DIMM, for τ_0, θ_0, and GLF observations are taken from MASS-DIMM. Model performance is given in Table 1. The density map of all the individual pairs of points is shown in Fig. D1.

vertical profiler having a vertical resolution $\Delta H \sim \sqrt{\lambda \left(h - h_{gs}\right)}/\theta$ (Avila, Vernin & Masciadri 1997). It therefore has a much higher resolution with respect to MASS. The vertical resolution is higher close to the ground and then it decreases with height and achieves an order of magnitude of $\sim$ 1 km in the free atmosphere (more than enough for most AO applications). The SS is basically a Generalised SCIDAR (GS), but with two independent cameras instead of one. This allows us to improve the signal-to-noise ratio (SNR) of the scintillation map cross-correlation and, as a consequence, the SS can provide a more precise reconstruction of the C_N^2 profile. Such a technical solution also allows us to increase the sky coverage, as the constraints on the Δm values of binary stars are relaxed with respect to the GS. The latter item is less critical because, in general, it is always possible to identify in the sky, during a whole night, a reasonable number of binaries respecting the constraints of the Generalised SCIDAR. The temporal sampling of measurements is not constant and is, on average, 2 minutes. The SS of ESO has been run on 157 nights distributed in different periods of different years between 2016 and 2019. The SS is implemented at the focus of one of the Auxiliary Telescopes (AT) having a diameter of 2.5 m. Stereo-SCIDAR measurements of the 2019B release (Butterley et al. 2020) have been corrected by the authors for several bugs identified in the previous release of 2018 (Osborn et al. 2018).

3 MODEL CONFIGURATION

The most recent version of the Astro-Meso-NH code (Masciadri et al. 1999a, 2017) has been used in this study for the forecast of optical turbulence, i.e. C_N^2 and four among the most relevant integrated astroclimatic parameters: seeing (ε), isoplanatic angle (θ_0), wavefront coherence time (τ_0), and GLF. The GLF is the portion of turbulence developed close to the ground with respect to that developed in the whole atmosphere (see Section 4 and equations 4 and 5). The Astro-Meso-NH is a package that has been developed on top of the MESO-NH[8] model (Lafore et al. 1998; Lac et al. 2018) that provides the spatio-temporal evolution of the hydrodynamic atmospheric flow. For this study we used, for the most part, the same geometrical model configuration used by Masciadri et al. (2013). The geographic coordinates of Cerro Paranal are (24°37′33.117″S, 70°24′11.642W) and the height of the summit is 2635 m above sea level. We used a grid-nesting technique (Stein et al. 2000) consisting of using different embedded domains of digital elevation models (DEM: i.e. orography) extended on smaller and smaller surfaces, with progressively higher horizontal resolution but with the same vertical grid. Simulations of OT are performed on three embedded domains where the horizontal resolution of the innermost domain is $\Delta X = 500$ m. In Appendix D we report information on the dimensions, number of grid points, and horizontal resolution of the three domains. Along the z-axis we have 62 levels distributed as follows: a first vertical grid point equal to 5 m, a logarithmic stretching of 20 per cent up to 3.5 km above the ground, and an almost constant vertical grid size of ∼600 m[9] up to 23.8 km. Regarding the MESO-NH model, the system of hydrodynamic equations is based

[8]The Meso-NH version used in this work is 5.2.

[9]Such a value is not exactly the same for all the (x,y) of a domain, as it depends on the orography. The model levels are indeed more similar to sheets than to perfect horizontal planes.

upon an anelastic formulation that allows an effective filtering of acoustic waves. The model uses the Gal-Chen & Sommerville (1975) coordinates system in the vertical and the C-grid in the formulation of Arakawa & Messinger (1976) for the spatial digitalization. In this study, in the wind advection scheme we used the 'forward-in-time' (FIT) numerical integrator instead of the 'leap-frog' one. Such a solution allows for longer time steps and therefore shorter computing time. The model employs a one-dimensional 1.5 turbulence closure scheme (Cuxart, Bougeault & Redelsperger 2000) and we used the one-dimensional mixing length proposed by Bougeault & Lacarrère (1989). The surface exchanges are computed using the Interaction Soil Biosphere Atmosphere (ISBA) module (Noilhan & Planton 1989). The model is initialized with forecasts provided by the GCM High Resolution Forecast (HRES) of the ECMWF, which has an intrinsic horizontal resolution of around 9 km. The forecast configuration used in this study is shown in Fig. 1(a). It corresponds to the operational configuration that we will use above VLT in the main baseline. The FTS is of the order of [23–33 h] depending on the period of the year, as the length of the night changes, and FTS_{disp} is of the order of [2–12 h]. Forecasts are delivered two hours before sunset (this corresponds to T_{disp}). We do not enter into details on why this configuration was selected, as it is not the topic of this work. It is enough to know that it depends on external requirements from ESO and might change in the future. We are interested in this work in using the main baseline operational configuration we have just described to quantify the forecast performance. For the short-time forecasts obtained with AR, which are the subject of this work, we used the same configuration used in Masciadri et al. (2020). The AR is calculated in post-processing and, in that case, the FTS is calculated with respect to the present time $t = 0$. In the definition of the FTS we have to replace $T_{fc} = 0$ and T_{ini} and T_{end} are, in this case, the extremes of the interval we are interested in (e.g. for the forecast at 2 h we have $T_{ini} = 1$ and $T_{end} = 2$ with respect to present time = 0. The forecasts at 2 h therefore include all values included in the [1 h, 2 h] range). When the night begins and real-time measurements start to be available, we perform the AR forecast extended to 4 h and, at each full hour, the forecast is updated for a successive 4 h. Starting from the instant at which real-time measurements are available, from these forecasts it is therefore possible to retrieve forecasts at 1 h, 2 h, 3 h etc. In conclusion, we can cover, for each night, forecasts with a great number of time-scales, starting from short ones with AR (i.e. 1 or 2 h) up to long ones obtained with the atmospheric model alone.

4 OT FORECAST PERFORMANCE

To optimize the AR performance we need contiguous nights (Masciadri et al. 2020) with available observations. To quantify the model performance, we therefore analysed simulations and observations related to a rich statistical sample of nights covering one full solar year [2018 August 1–2019 July 31].

We considered observations coming from DIMM for the seeing, and observations coming from MASS-DIMM for τ_0, θ_0, and GLF. The reason for this choice is that ESO identified, for each parameter, a specific instrument of reference.[10]

[10]We refer here to ESO CFT/ESO/20/95952/FLAB Call for tender that refers to the forecast of optical turbulence above the VLT. These are therefore technical specifications we have to respect and this also means that, when we calculate forecast performance in the standard configuration, we also use these references. With the AR approach we intend to improve the forecast performance on short time-scales using in situ measurements following ESO's desiderata.

Table 1. Statistical operators (BIAS, RMSE, SD) related to Fig. 3 (or Fig. D1) obtained with observations and the AR method on a time-scale of 1 h. Selection for seeing ≤ 1.5 arcsec is done on observations.

	ε (arcsec)	$\varepsilon \leq 1.5$ arcsec (arcsec)	τ_0 (ms)	θ_0 (arcsec)	GLF × 100 (%)
BIAS	0.00	0.00	− 0.01	− 0.01	0.00
RMSE	0.10	0.09	0.58	0.16	0.06
SD	0.10	0.09	0.58	0.16	0.06

Fig. 3 shows the scatter plot obtained by comparing observations and AR forecasts calculated on a time-scale of 1 h for the seeing ε, wavefront coherence time τ_0, isoplanatic angle θ_0, and GLF. For each parameter we considered observations from the associated instruments of reference.

The original temporal frequency of the forecast is 2 minutes. Forecasts and observations have been treated with a moving average of 1 h followed by resampling on a time-scale of 10 minutes. Precedent studies (Lascaux, Masciadri & Fini 2015; Masciadri et al. 2017) showed that such a treatment was the most suitable for this kind of application/analysis. We refer the reader to Masciadri et al. (2020) for further details related to the AR technicalities. Table 1 reports values of BIAS, RMSE, and SD for all the astroclimatic parameters related to a FTS equal to 1 h. BIAS, RMSE, and SD are calculated as

$$BIAS = \sum_{i=1}^{N} \frac{Y_i - X_i}{N}, \tag{1}$$

$$RMSE = \sqrt{\sum_{i=1}^{N} \frac{(Y_i - X_i)^2}{N}}, \tag{2}$$

$$SD = \sqrt{\sum_{i=1}^{N} \frac{[(X_i - Y_i) - (\overline{X_i - Y_i})]^2}{N}} = \sqrt{RMSE^2 - BIAS^2}, \tag{3}$$

where X_i and Y_i are the observations and AR method respectively, N is the number of times for which pairs (X_i, Y_i) are available and both X_i and Y_i are different from zero. For the seeing we calculated the case in which we took into account all the seeing values and the case in which the observed seeing ≤ 1.5 arcsec. As can be seen in Table 1, the statistical operators for the seeing in the two cases (column 2 and 3) are very similar. This tells us that, using the AR method, the forecast performance is very good for whatever threshold is on the seeing value. For long-time-scale forecasts, the forecast performance for the seeing decreases when the seeing increases. In that case we therefore have interest in calculating the forecast performance only in the range $\varepsilon \leq 1.5$ arcsec where AO is effective. In Fig. D1 we show the density function maps associated with Fig. 3 that indicate better where the peak of the distribution of the (X_i,Y_i) pairs falls. We observe that, for all four parameters, the peak falls on the bisector, telling us that the BIAS is basically null. The values of the statistical operators (Table 1) indicates a RMSE of the order of 0.09 arcsec for the seeing, 0.58 ms for τ_0, 0.16 arcsec for θ_0, and 6 per cent for GLF.

As mentioned in appendix A of Masciadri et al. (2020), if it is not possible to collect observations related to the previous five nights (for example because there are no measurements) then AR is not executed. The temporal coverage of observations is very good at the VLT. For this reason, for the four astroclimatic parameters analysed in this study the number of missed forecasts with AR is very small. We counted only 17 missed nights for the seeing and 18 for the other astroclimatic parameters in the solar year tested in this work. The number of nights N considered in the past (see Masciadri

Table 2. Accuracy obtained with observations: Stereo-SCIDAR (SS) and MASS-DIMM. First row: $BIAS_{obs}$ calculated by comparing simultaneous values from different instruments calculated on a rich statistical sample. Second row: as the first row but $RMSE_{obs}$ is shown. Third row: as the first row but standard deviation SD_{obs} is shown.

Param.	Seeing (ε) (arcsec)	Wavefront coherence time (τ_0) (ms)	Isoplanatic angle (θ_0) (arcsec)	GLF ($\times$ 100 %)
Instruments	SS vs. DIMM	SS vs. MASS-DIMM	SS vs. MASS-DIMM	SS vs. MASS-DIMM
No. nights	157	142	142	142
$BIAS_{obs}$	0.08	0.41	− 0.16	0.15
$RMSE_{obs}$	0.25	1.29	0.45	0.20
SD_{obs}	0.24	1.22	0.42	0.14

et al. 2020) for the application of AR is a trade-off between the temporal coverage of the technics and the performance one wants to obtain. If N decreases, we have a larger temporal coverage but slightly lower performance (see for example fig. 7 of Masciadri et al. 2020). That means that the number of missed nights might be smaller if we accept decreased model performance. These are technicalities to be considered in the management of the individual observatories.

How can we judge the goodness of the forecast performance? What we want to show here is that, if the RMSE is smaller than the intrinsic achievable accuracy in estimating the OT, we can consider that the achieved RMSE fits with good and realistic expectations.

To answer to this question, we therefore compare the RMSE that we obtained with our analysis with the 'typical' dispersion of estimates obtained with different and independent instruments conceived for turbulence estimates running simultaneously calculated on a rich statistical sample of nights. The standard deviation (SD) between the observations from different instruments provides us with the 'typical' uncertainty, i.e. the accuracy with which we can reasonably estimate these parameters with observations. If we look at the SD, that means we are assuming we eliminate whatever systematic errors exist due to bias/error in the instruments and we are looking at the intrinsic statistical random error that has to be taken into account. This is a very important passage as, even if one elects for some reason a preferred/reference instrument with respect to another one, we cannot forget that the ultimate accuracy is given by the SD of measurements obtained with different instruments. The calculation of the typical uncertainty (SD) related to instruments is very useful because it allows us to have a term of reference. To calculate the typical dispersion (or uncertainty), we selected a statistical sample of nights in which Stereo-SCIDAR and MASS-DIMM (or simple DIMM depending on the parameter we are considering) were run simultaneously at Cerro Paranal. Such a sample of 157 nights has two important characteristics. (1) It is one of the richest (in statistical terms) available samples obtained with SCIDAR and MASS-DIMM at present in the astronomical context. Considering the very large number of nights, we can assume that the sample is statistically representative of the general conditions of turbulence. (2) Even if in principle the goal here is to quantify a typical intrinsic SD for a typical good astronomical site, the fact that these measurements have been done at Cerro Paranal is particularly attractive for us, as this is the same site at which our study is performed.

Table 2 reports the BIAS, RMSE, and SD between observations of the four astroclimatic parameters obtained with Stereo-SCIDAR and MASS-DIMM (see Section 2). Following equations (1)–(3), here X_i are the individual values of Stereo-SCIDAR and Y_i the observations from MASS-DIMM. DIMM and MASS-DIMM are the references used by ESO as described previously. As described previously, DIMM is used for the seeing and MASS-DIMM for τ_0, θ_0, and GLF.

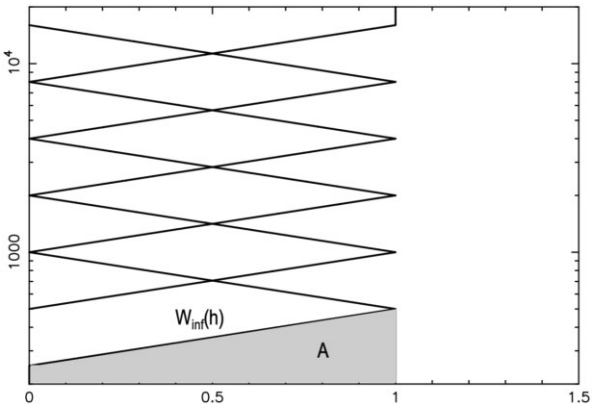

Figure 4. Weighting functions of the MASS given by typical triangle functions. The grey zone indicated with 'A' is the numerator of equation (4).

In the case of the seeing, the number of nights on which we have observations from both Stereo-SCIDAR and DIMM is the whole data set of 157 nights. For τ_0, θ_0, and GLF, we have observations from Stereo-SCIDAR and MASS-DIMM on a slightly smaller number of nights: 142 nights. The GLF retrieved from Stereo-SCIDAR (to be compared with that calculated from MASS-DIMM) has been calculated according to the following procedure: we considered the C_N^2 profile from Stereo-SCIDAR projected on the MASS weighting functions $W_i(h)$ (having a typical triangle shape – see Fig. 4) and we took the complementary part close to the ground (not sensed by MASS), which corresponds to a trapezoid (A in Fig. 4). This trapezoid corresponds to the numerator in the definition of the GLF of MASS-DIMM (see equation 4):

$$GLF = A/J, \tag{4}$$

where J is the integral of C_N^2 over the whole atmosphere as provided by MASS-DIMM. The oblique line $W_{\rm inf}(h)$ in Fig. 4 is the lower limit of the MASS weighting function. It is given by

$$W_{inf}(h) = \begin{cases} 0 & \text{if } h \leq 250\ \text{m}, \\ 1 - \left|1 - \log_{10}\left(\frac{h}{250}\right) \cdot \frac{1}{\log_{10} 2}\right| & \text{if } 250\ \text{m} < h < 500\ \text{m}, \\ 1 & \text{if } h \geq 500\ \text{m}. \end{cases} \tag{5}$$

By comparing SD_{obs} (Table 2) with the RMSE of the AR method (Table 1), we observe that RMSE is always $\leq SD_{obs}$ for all four astroclimatic parameters, and this allows us to conclude that the forecast performance of the AR method at 1 h can be considered very satisfactory, as we obtain a dispersion/error that is of the same order as the uncertainty obtained with instruments and, in some cases, even better. As anticipated previously, we compare RMSE with SD_{obs} to make the challenge even more difficult for the forecasts.

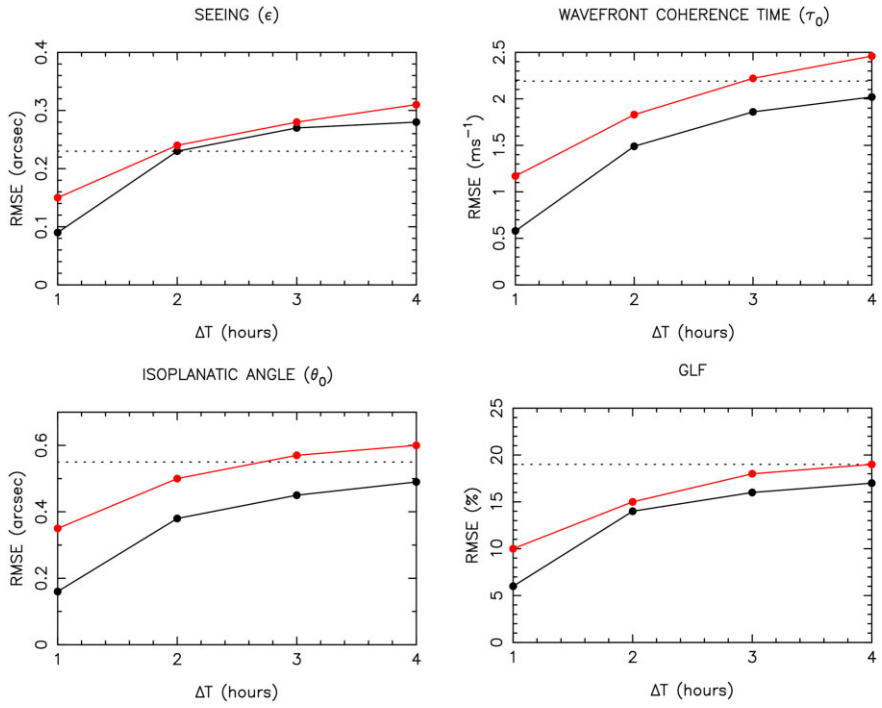

Figure 5. RMSE vs different FTS ΔT for seeing, wavefront coherence time, isoplanatic angle, and GLF. On the x-axes the 'forecast time' ΔT as defined in Section 3. The black dashed line is the RMSE related to the standard forecast, the black line the AR method, and the red line the forecast by the persistent method. Seeing is calculated taking all observed values ≤ 1.5 arcsec.

We indeed assume a null BIAS_{obs} (i.e. no systematic errors in instrumentation) even if it is not necessarily the case now, as is visible in Table 2. In other words, we compare performance of the forecast system with the pure intrinsic statistical uncertainty provided by observations.

What about the forecast performance of the AR method at FTS different from 1 h?

Fig. 5 shows the RMSE for different FTS (black line) compared with the RMSE for the FTS of the 'standard forecast' (black dashed line). On time-scales of 1 and 2 h, we observe that the AR method provides better results than the 'standard forecast' (i.e. the RMSE has a smaller value). This means that the AR is very effective on short time-scales, as expected. We observe that, while for the seeing the gain with respect to the 'standard forecast' disappears after around 2 h (as the black line intersects the black dashed line), for the other astroclimatic parameters the gain still survives up to 4 h or more. That said, by repeating the AR forecast with a sampling of 1 h as explained in Masciadri et al. (2020), it is possible always to guarantee the AR forecasts on time-scales of 1 and 2 h all through the night.

We now compare the AR method with a very simple and independent benchmark such as the method of forecast by persistence.

Performance of the forecast by persistence is represented in Fig. 5 by a red line. The method of forecast by persistence in this context assumes that the forecast extended to the successive 4 h with respect to t_0 is equal to the value of the observed parameter at time t_0. Looking at Fig. 5, we observe that the AR method provides better performance than the method of forecast by persistence for all four astroclimatic parameters at all time-scales even if the gain of the AR with respect to the persistence method is not the same for the different parameters. Table 3 reports the gain of the AR with respect to the 'standard forecast' and the forecast by persistence at 1 h for all four parameters.

We conclude that the AR method is definitely more performant than the 'standard forecast' as well as the forecast by persistence for all parameters. The 'standard forecast' is more performant than the AR method and the forecast by persistence at long time-scales when the beneficial effects of the AR approach disappear. Fig. 6 shows the relation between the method by persistence and the forecasts that we deliver (AR method and standard configuration respectively for the short and long time forecasts) for the seeing. For $\Delta T \leq 2$ h, the AR method is better than prediction by persistence. For $\Delta T > 2$ h, when the positive effect of AR decreases, the standard configuration takes the relay and, in this range, provides better performance than the prediction by persistence. By definition, the latter loses effectiveness when ΔT increases. This means that up to $\Delta T = 2$ h it is convenient for the seeing to look at AR and for $\Delta T > 2$ h it is preferable to consider the forecast at long time-scale obtained with the standard configuration.

Table 3. Gain of AR at 1 h with respect to the 'standard forecast' and the prediction by persistence. First row: gain for the RMSE for different astroclimatic parameters (seeing, wavefront coherence time, isoplanatic angle, and ground-layer fraction) of AR with respect to the 'standard forecast' i.e. the RMSE obtained for the standard configuration divided by that obtained for AR. Second row: gain of the AR method with respect to prevision by persistence on the same time-scale, i.e. the RMSE obtained for the PP divided by that obtained for AR. Third row: gain of AR versus prevision by persistence expressed in per cent, i.e. the value of the second row -1 expressed in per cent. Seeing is calculated taking all observed values ≤ 1.5 arcsec.

GAIN	ε	τ_0	θ_0	GLF
AR vs. SC	2.55	3.77	3.43	3.16
AR vs. PP	1.67	2.02	2.19	1.67
%	67	102	119	67

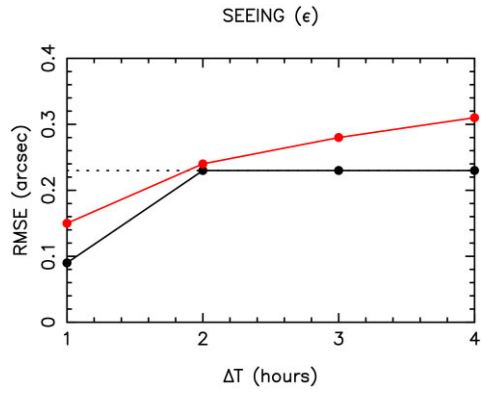

Figure 6. As Fig. 5, but here the black line (bold style) represents the best configuration of our forecast system to be compared with prediction by persistence. For $\Delta T \leq 2$ h we have to take into account the AR method. For $\Delta T > 2$ h the configuration providing the best gain with respect to the prediction of persistence is the forecast at long time-scale, i.e. what we call the 'standard configuration' (see text).

4.1 OT forecast performance on individual nights

In the previous section we analysed the dispersion between observations (DIMM and MASS/DIMM depending on the parameter) and AR forecasts with statistical operators applied to the whole sample of (X_i, Y_i) pairs taken all together (see Fig. 3 and Table 1). In this section the AR method forecast performance is analysed following a different approach that we call *'on individual nights'*. We analyse the statistical operators BIAS, RMSE, and SD on each night and then we calculate the cumulative distribution and derive the median and the quantiles. In precedent studies (Lascaux et al. 2015) it has already been observed that the forecast performance improves slightly if analysed following this approach. This is not surprising, as in the second case the statistical operators are calculated first on a smaller sample, i.e. the individual nights, while in the first case the statistical operators are calculated on the whole sample of data. The use of one or the other method depends on which feature/characteristics of the system one wishes to put in evidence. The important thing is, however, to declare which method/approach has been used, as one method provides intrinsically better estimates than the other. Fig. 7 shows the cumulative distribution of the RMSE of all astroclimatic parameters (seeing ε, isoplanatic angle θ_0, wavefront coherence time τ_0, and ground-layer fraction GLF) in two cases: the short-time-scale forecasts at 1 and 2 h. Table 4 reports the corresponding median, first, and third quartiles for all astroclimatic parameters for forecasts calculated at 1 h (first row) and 2 h (second row).

We observe that, for the forecasts at 1 h, the median values of the RMSE are 0.08 arcsec for the seeing, 0.44 ms for the wavefront coherence time τ_0, 0.14 arcsec for the isoplanatic angle θ_0, and 6 per cent for the GLF. The dispersion is very narrow, indeed the third quartiles at 1 h still remain smaller than the RMSE obtained with different instruments (see Table 2). The median values of the RMSE at 2 h (i.e. 0.19 arcsec for the seeing, 1.01 ms for τ_0, 0.32 arcsec for θ_0, and 13.72 per cent for GLF) slightly increase with respect to the 1-h case but still remain smaller than the typical dispersion of instruments (SD_{obs}). These results confirm the conclusions obtained in Section 4, i.e. that the forecast performance obtained with this 'beta' configuration is very promising.

4.2 Probability of detection for short FTS using the AR method

To complete the analysis, we calculate here the contingency tables and associated statistical operators, i.e. the probability of detection (POD), percentage of detection (PC), and extremely bad detection (EBD). Contingency tables allow for the analysis of the relationship between two or more categorical variables. We refer the reader to Lascaux et al. (2015) for a description of this method. In Appendix B we report a synthesis of the definition of the different operators used in this work to quantify forecast performance: contingency tables, POD, PC, and EBD. Given a statistical sample of observations and predictions, the contingency tables allow us to calculate the number of times at which observations and forecasts fall in the same intervals of values. For all four astroclimatic parameters treated in this work (seeing ε, wavefront coherence time τ_0, isoplanatic angle θ_0, and ground-layer fraction GLF), we considered 3×3 contingency tables in which the thresholds separating the different categories are retrieved from the climatological estimates. More precisely, we used as thresholds for the seeing and the GLF the first and second tertiles of the distribution calculated on the same sample of observations (a whole solar year) used for the analysis of the AR method. For τ_0 and θ_0, the thresholds are the second and first tertiles (see Table 5).[11] From the contingency tables, we retrieve the probability of detection POD of the parameter X within a specific range of values.

To quantify the POD, we take into account the accuracy, which allows us to say in which category of the generic table 3×3 the estimates belong. To estimate the accuracy (hereafter ACC), we take ACC equal to the standard deviation (SD_{obs}) obtained by comparing different instruments running simultaneously. In other words, we assume no biases between the instruments and we consider only the dispersion due to the statistical uncertainty. The ACC values for the four astroclimatic parameters are therefore imposed as equal to the SD_{obs} reported in Table 2. What we show here is that, with the AR method, we can obtain POD_i that are, in most cases, well above 90 per cent for realistic levels of ACC in the case of forecasts at 1 and 2 h.

Tables 6–9 report results of POD_i, PC, and EBD obtained for each astroclimatic parameter at short time-scales with the AR method i.e. at (a) 1 h and (b) 2 h.

The best performance is achieved, as expected, in case (a), where all POD_i are of the order of 99 per cent for all four astroclimatic parameters.

Performance decreases only slightly at a 2-h time-scale, i.e. in case (b). In most cases, it remains above 90 per cent. We highlight that this is true in particular in the most interesting categories from an astronomical point of view, i.e. the first tertile (POD_1) for the seeing and POD_3 for τ_0 and θ_0. Indeed, for the longest period of 2 h we have $POD_1 = 92$ per cent for the seeing, $POD_3 = 90$ per cent for τ_0, and $POD_3 = 91$ per cent for θ_0.

For the GLF we have very good performance at 1 h (all POD_i well above 98 per cent); forecast performance decreases but still remains satisfactory at 2 h, with POD_1 and POD_2 above 90 per cent and POD_3 slightly smaller (83.53 per cent).

It is worth adding a few comments on the latter parameter. The GLF is sensitive to the boundary layer as well as the total turbulence developed in the atmosphere, therefore uncertainties in the numerator and denominator both increase the total uncertainty. The threshold separating the boundary from the free atmosphere (W_{inf};

[11] For τ_0 and θ_0, the cumulative distribution from which the median and tertiles are retrieved is calculated from monotonically decreasing values.

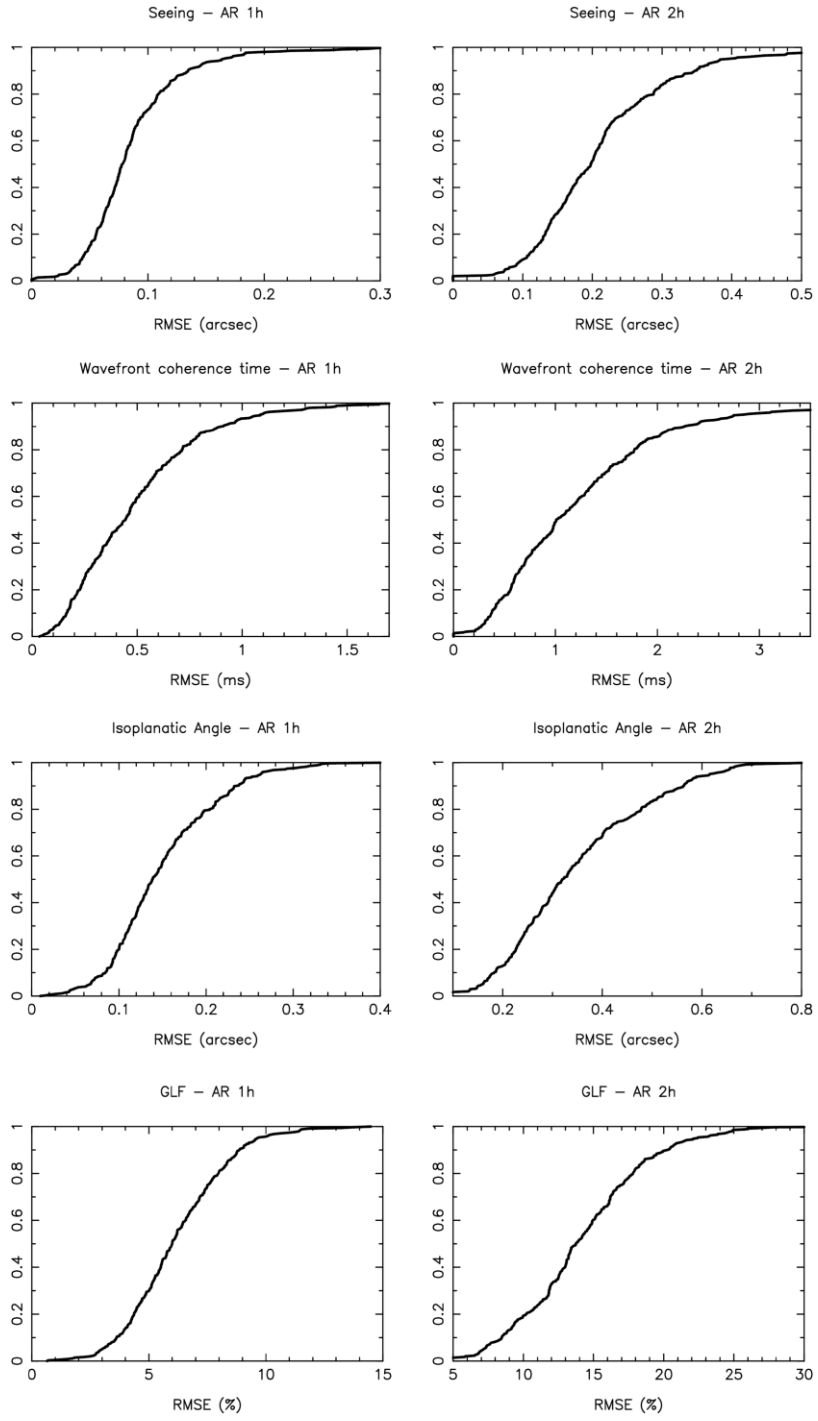

Figure 7. Cumulative distribution of the RMSE calculated between observations and forecasts related to the astroclimatic parameters (ε, τ_0, θ_0, and GLF) obtained following the 'individual nights' approach on a sample of 1 solar year. Seeing is calculated with the condition that observed values are ≤ 1.5 arcsec; see Section 4. First column: AR forecasts at 1 h. Second column: AR forecasts at 2 h. See Table 4 for the quantitative estimates of median, first, and third quartiles. The scale on the x-axis has been selected to optimize the dynamic of each figure.

see Fig. 4) falls more or less in the region where the sensitivity of the atmospheric model to the OT parametrization changes, therefore this makes things more difficult. Moreover, we also observe that the GLF is the sole parameter for which there is still a non-negligible bias between Stereo-SCIDAR and MASS-DIMM (see Table 2), with a larger GLF seen by MASS-DIMM. Such an effect has already been observed in different contexts/studies. There is not a unanimous consensus on the origin of this discrepancy. Some authors think that it is due to uneven conditions extending on the Cerro Paranal plateau close to the surface that determine inhomogeneous GLF estimates made by instruments located at a certain distance with respect to one another on the plateau (Haguenauer, Giuesalaga & Butterley 2020). However it might also be possible that residual systematic biases in the instrumentation are still present, or perhaps

Table 4. Median, first, and third quartiles of the RMSE calculated for different astroclimatic parameters on individual nights: seeing (ε), wavefront coherence time (τ_0), isoplanatic angle (θ_0), and ground-layer fraction GLF (see associated cumulative distribution in Fig. 7). Seeing is calculated taking all observed values ≤ 1.5 arcsec. Sub/superscripts refer to the first and third quartiles.

RMSE	ε (arcsec)	τ_0 (ms)	θ_0 (arcsec)	GLF (%)
@ 1 h	$0.08^{0.10}_{0.06}$	$0.44^{0.65}_{0.25}$	$0.14^{0.19}_{0.11}$	$6.00^{7.50}_{4.65}$
@ 2 h	$0.19^{0.26}_{0.14}$	$1.01^{1.65}_{0.60}$	$0.32^{0.44}_{0.24}$	$13.72^{17.00}_{11.20}$

Table 5. Median, first, and second tertiles of the distribution of the four astroclimatic parameters (ε, τ_0, θ_0, and GLF) calculated for a sample of observations related to one full solar year (always the same [2018 August 1–2019 July 31]). Measurements of the seeing have been obtained with DIMM, those of the other astroclimatic parameters with MASS-DIMM (see Section 4).

Parameter	Median value	First tertile	Second tertile
ε (arcsec)	0.68	0.59	0.80
τ_0 (ms)	4.46	5.79	3.37
θ_0 (arcsec)	1.92	2.22	1.65
GLF ($\times$ 100 %)	0.64	0.56	0.71

Table 6. Probability of detection (POD_i) with percentage of correct detection (PC) and extremely bad detection (EBD) as retrieved from the contingency tables (see Appendix B) calculated for the seeing (ε) in two different conditions: (a) forecast at 1-h time-scale with AR and (b) forecast at 2-h time-scale with AR. Thresholds of the 3×3 tables are the first and second tertiles calculated using a sample of measurements taken over one solar year (see Table 5). Values of POD, PC, and EDB are calculated assuming an accuracy ACC $= SD_{obs} = 0.24$ arcsec (see Table 2). We considered seeing $<$ 1.5 arcsec.

Param.	Seeing (ε) Forecast with AR at 1 h (%)	at 2 h (%)
POD_1	99.80	94.29
POD_2	99.42	91.43
POD_3	99.35	89.33
PC	99.62	93.19
EBD	0.19	3.95

both of them play a role. In conclusion, the estimate of the GLF with instruments is not yet a closed problem. Even if there is still space for further improvements on the side of modelling, discrepancies in observed GLF from different instruments tell us that this parameter, as it is measured at the present time, deserves further investigation for use as a robust figure of merit, even if the portion of turbulence close to the ground with respect to the total turbulence is in principle very useful to quantify atmospheric conditions more or less favourable for observations supported by WFAO. We do not comment further, to avoid going outside the scope of this work, but work is in progress to improve GLF predictions.

Anyway, in cases (a) and (b) the EBD values are extremely low for all astroclimatic parameters, with a maximum of ~ 5 per cent, and this guarantees us a very low probability of failing the forecasts.

Table 7. As Table 6 for the wavefront coherence time (τ_0), but the thresholds are the second and first tertiles (see Table 5). Values calculated assuming an accuracy ACC $= SD_{obs} = 1.22$ ms (see Table 2).

Param.	Wavefront coherence time (τ_0) Forecast with AR at 1 h (%)	at 2 h (%)
POD_1	99.86	97.16
POD_2	99.51	94.30
POD_3	98.86	89.76
PC	99.60	95.32
EBD	0.20	1.96

Table 8. As Table 6 but for the isoplanatic angle (θ_0), but the thresholds are the second and first tertiles of the distribution (see Table 5). Values calculated assuming an accuracy ACC $= SD_{obs} = 0.42$ arcsec (see Table 2).

Param.	Isoplanatic angle (θ_0) Forecast with AR at 1 h (%)	at 2 h (%)
POD_1	99.84	96.41
POD_2	99.88	95.24
POD_3	99.27	91.34
PC	99.78	99.15
EBD	0.12	1.81

Table 9. As Table 6 but for the ground-layer fraction GLF (see Table 5). Values are calculated assuming an accuracy ACC $= SD_{obs} = 14\%$ (see Table 2).

Param.	Ground-layer fraction (GLF) Forecast with AR at 1 h (%)	at 2 h (%)
POD_1	99.13	92.05
POD_2	98.91	91.44
POD_3	98.77	83.53
PC	99.14	91.34
EBD	0.44	5.47

5 MACHINE LEARNING VERSUS AR

In this section we present an analysis with the goal of quantifying the performance obtained in predicting the astroclimatic parameters described in the previous sections at short time-scale with the random-forest (RF) machine-learning approach using only observations. The idea behind the machine-learning approach is to make use of past observations of N parameters to estimate the parameter X (depending on the other N parameters) in the very close future (1 and 2h). The RF approach has been previously applied to the seeing by Milli et al. (2019). That work showed performance of the predictions in the close future that is no better than the prediction by persistence method. The goal of our exercise is to reproduce the experiment and extend such a calculation to all other astroclimatic parameters. Even more importantly, we need to treat the outputs in a homogenous way[12] with respect to AR and the persistence method to be able to compare the three approaches.

The RF (and in general whatever ML algorithm we consider) implies a twofold step to forecast the parameter X: a 'training' of the algorithm performed on a rich statistical sample of observations of several parameters related to X (in our case meteorological and

[12] As described in Section 3, we are considering steps of one hour.

astroclimatic parameters) and a 'test' of the system that consists of applying the trained algorithm on an independent data set to provide the forecast parameter X without any previous knowledge of the parameter to be predicted.

We do not enter into details of the RF configuration, as a dedicated forthcoming article focused on this topic is in preparation. The article will also treat other machine-learning and deep-learning approaches. We report here just the results of the RF for the four astroclimatic parameters and discuss them with respect to AR and prediction by persistence. We mention only that we used the same conditions used by Milli et al. (2019) and the RF as implemented in the PYTHON packages scikit-learn (Pedregosa et al. 2011).

As done by Milli et al. (2019), we used as inputs of the ML algorithm seeing, τ_0, GLF, wind speed (WS) at 200 mb, PWV, pressure at the ground (P), temperature (T), WS, and relative humidity (RH) at 30 m. To take into account the seasonal and daily variability trends of the values of the parameters, we included in the data set the following variables: cos(day/365), sin(day/365), cos(hour/24), sin(hour/24). We estimated the optimal length of the training data set by varying the selected sample length and observing at which point the performance of the algorithm no longer increases on adding more data (saturation level). This is in general between 1 and 1.5 years. The 'test' has been obtained with an independent data set relative to a 0.5-year range. The input data have been resampled to have one data point every 5 min. We expressly used the same conditions used by Milli et al. (2019). We simply remember that the prediction of the parameter X on a 1-h time-scale can be performed if there are observations of the parameter X (and the N related parameters) during the previous 1 h, and likewise for the prediction at a 2 h time-scale. Fig. 8 shows the outputs (blue dots) related to the RF approach obtained at 1 and 2 h, compared with the AR and prediction by persistence approaches. From Fig. 8 we conclude that, in all cases, performance obtained with the RF machine-learning technique provides worse results than both the AR approach and the prediction by persistence method. RF therefore seems non-competitive even with respect to prediction by persistence. We limit the analysis to the first short periods of 1 and 2 h, which are the most relevant and interesting ones. We performed some tests on these calculations to quantify (1) how the period of training being taken at different times in the past impacts on the estimates and (2) what is the impact of the length of the 'test' sample on results. For case (1), we observed a variability of the order of +/− 10 per cent. Treating case (2), we observed that, on using 'test' samples from 6 months up to 2 years, the RMSE increases by up to +10 per cent in the 1 h case and +15 per cent in the 2 h cases. For the reason we have just explained, in case (2) the RMSE can become larger if we consider longer 'test' samples. We conclude therefore that the values reported in Fig. 8 are quite conservative, and therefore the comparison of the ML, prediction by persistence, and AR methods is very robust.

Our next step is to study the use of other ML algorithms and/or ML algorithms joint to the numerical prediction provided by atmospheric models to investigate whether it is possible to obtain further improvements with respect to the AR approach that we consider, at the present time, as our benchmark threshold.

6 CONCLUSIONS

In this work we demonstrate that, with the Astro-Meso-NH model, configured as described in this work, using the techniques of autoregression (AR) proposed in Masciadri et al. (2020), we can provide forecasts at short time-scales (i.e. up to 1 and 2 h) of the four main astroclimatic parameters (ε, τ_0, θ_0, GLF) with an RMSE that is consistently smaller than or equal to the instrument accuracy. Perspectives are therefore very satisfactory to support the service mode of the Very Large Telescope, even more if we consider that our operational forecast system is just a beta version and that a set of modifications that will probably lead to an improvement of results is yet to be implemented in the system (see Section 1).

Quantitative estimates of the model performance are obtained following three different approaches. First we provide the BIAS, RMSE, and SD calculated following two different methods: (1) by considering the whole data set extended over one solar year as a unique sample; (2) by calculating the statistical operators (BIAS, RMSE, SD) on the individual nights and then retrieving the statistics of these results (median and quantiles) for a sample of one solar year. We then provide (3) the probability of detection (POD_i), percentage of detection (PC), and probability of extremely bad detection (EBD) as retrieved from the contingency tables. Quantitative estimates have been performed on a sample of one solar year of independent nights.

Following method (2) we obtain the best RMSE, as expected. On a time-scale of 1 h we have the following median values for the RMSE: 0.08 arcsec for the seeing, 0.44 ms for τ_0, 0.14 arcsec for θ_0, and 6 per cent for the GLF. On a time-scale of 2 h, the median RMSE is 0.19 arcsec for the seeing, 1.01 ms for τ_0, 0.32 arcsec for θ_0, and 13.72 per cent for GLF. By comparing the RMSE between forecasts and observations obtained in our analysis with the accuracy STD_{obs} related to observations from different instruments (currently used by ESO above Cerro Paranal), we concluded that RMSE $\leq$ STD_{obs} for all four astroclimatic parameters up to a time-scale of 2 h. This guarantees us robust performance of the forecast system for the service mode.

In this study we also demonstrate that, for all four astroclimatic parameters (seeing, wavefront coherence time, isoplanatic angle, and GLF) the forecasts at 1 and 2 h obtained using the AR method (atmospheric model plus real-time observations) provide a substantial smaller RMSE with respect to the simple use of an atmospheric model, i.e. the 'standard configuration'. The gain of AR with respect to the standard configuration is of the order of $\sim$ [2.6–3.8], depending on the parameter, on time-scales of 1 h. The gain of AR decreases when the FTS increases and disappears at different ΔT_i depending on the astroclimatic parameter (see Fig. 5). The shortest ΔT at which the gain disappears is around 2 h for the seeing; for all other astroclimatic parameters the gain persists for longer time-scales.

We demonstrate also that the AR technique provides better performance than prediction by persistence for all four astroclimatic parameters, with gains that depend on the parameter: they are included in the range [1.7–2.2] (corresponding to [67–119 per cent] – see Table 3) on a time-scale of 1 h.

Concerning the contingency tables and the probability of detection (POD_i), we calculated, for the four astroclimatic parameters, that, taking into account the intrinsic uncertainty related to the turbulence estimates, on a time-scale of 1 h all POD_i are of the order of 99 per cent. Performance decreases only slightly on a time-scale of 2 h, but in most cases remains above 90 per cent. This is true in particular for the most interesting categories from an astronomical point of view, i.e the probability to detect a value below the first tertile (POD_1) for the seeing and the probability to detect a value above the third tertile (POD_3) for τ_0 and θ_0. On a 2-h time-scale, we have $POD_1 = 94.29$ per cent for the seeing, $POD_3 = 89.76$ per cent for τ_0, and $POD_3 = 91.34$ per cent for θ_0. We observe a slightly smaller value $POD_3 = 83.53$ per cent for GLF at 2 h. As discussed in this work, this slightly poorer performance is probably due to a set of reasons, not necessarily related to the forecast method (see Section 4.2).

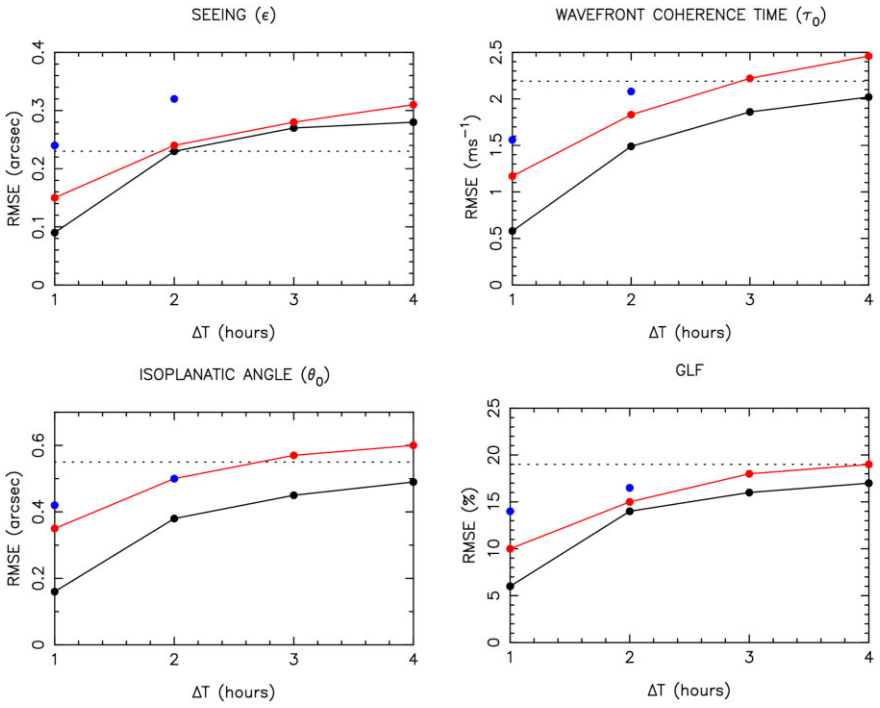

Figure 8. As Fig. 5 but with the addition of machine learning (Random Forest) outputs (blue dots). Black line: AR method. Red line: prediction by persistence.

Further analyses are planned in the future. For example, we might calculate the POD_i for smaller values of ACC. We estimated that, with ACC = SD_{obs} = 0.1 arcsec (instead of 0.24 arcsec), all the POD_i for the seeing at 1 h are in any case larger than or equal to 95 per cent. This tells us that, even in the case of a smaller SD_{obs}, forecast performance at short time-scales is very promising.

Finally, we compared our results with a 'random forest' (RF) machine-learning algorithm. We observe that, on time-scales of 1 and 2 h, the AR method provides better results with respect to the RF. The latter appears, in most cases, even worse than prediction by persistence.

We definitely think that we are entering into a new era of hybrid techniques made up of numerical techniques (i.e. atmospheric models) plus statistical techniques that aim to improve forecast performance. The latter include a set of various approaches such as the autoregression method, Kalman filter, machine learning, deep learning, artificial neural networking (ANN), etc. We showed that machine learning alone, performed with just observations without the support of an atmospheric model, is not as good as hybrid techniques, therefore atmospheric models still play a crucial role.

In terms of perspectives, we intend to push the use of machine learning and deep learning further, with the goal of improving the performance of the operational forecast system that we developed. This can be done by testing different and more complex algorithms or using more complex configuration models/ML. Studies are in progress. An automatic operational version of the system presented in this work will be implemented soon at the VLT (FATE project).

ACKNOWLEDGEMENTS

The authors thank the Meso-NH user support team, who work constantly to maintain the model by developing new packages in progressing model versions. Initialization data come from the GCM HRES of the ECMWF. This study has been co-funded by the FRCF foundation through the 'Ricerca Scientifica e Tecnologica' action – N.45103 and by the EU Horizon 2020 research and innovation programme under grant agreement No. 824135 (SOLARNET). The digital elevation model at high resolution has been obtained thanks to the Shuttle Radar Topography Mission (SRTM). The authors thank Angel Otarola and ESO Santiago and Garching staff supporting this study. This study made use of the scikit-learn PYTHON packages. We thank Leslie Hunts for her precious hints on the English form of the text.

DATA AVAILABILITY

Model initialization data come from the ECMWF atmospheric general forecast model HRES. Measurements from MASS-DIMM are part of the ESO archive.[13] Data from Stereo-SCIDAR are part of Release 2019B – Durham University (Butterley et al. 2020). Data obtained with model simulations are not shared publicly because of intellectual property right. They can be shared on reasonable request to the corresponding author.

REFERENCES

Arakawa A., Messinger F., 1976, GARP Tech. Rep., 17. WMO/ICSU, Geneva, Switzerland
Avila R., Vernin J., Masciadri E., 1997, Appl. Optics, 36, 7898
Basu S., Osborn J., He P., De Marco W., 2020, MNRAS, 497, 2302
Bolbasova L. A., Andrakhanov A. A., Shikhovtsev A. Yu., 2021, MNRAS, 504, 6008
Bougeault P., Lacarrère P., 1989, Mon. Weather Rev., 117, 1972
Butterley T., Sarazin M., LeLouarn M., Osborn J., Farley O., 2020, SPIE, 11448, 114481W
Cantalloube F. et al., 2018, A&A, 620, L10
Cantalloube F. et al., 2020, A&A, 638, A98
Cherubini T., Businger S., 2013, J. Appl. Meteorol. Climatol., 52, 498

[13] http://archive.eso.org/cms/eso-data/ambient-conditions.html

Cherubini T., Businger S., Lymann R., Chun M., 2008, J. Appl. Meteorol. Climatol., 47, 1140
Cherubini T., Lyman R., Businger S., 2022, MNRAS, 509, 232
Coulman C. E., André J. C., Lacarrére P., Gillingham P. R., 1986, PASP, 98, 376
Cuxart J., Bougeault P., Redelsperger J.-L., 2000, Q. J. R. Meteorol. Soc., 126, 1
Dewan E., Good R., Beland R., Brown J., 1993, Profiles Using Radiosonde Data, PL-TR-93-2043. NTIS, Springfield, VA
Gal-Chen T., Sommerville C. J., 1975, J. Comput. Phys., 17, 209
Giordano C., Vernin J., Vazquez Ramio H., Munoz-Tunon C., Varela A. M., Trinquet H., 2013, MNRAS, 430, 3102
Hagelin S., Masciadri E., Lascaux F., 2011, MNRAS, 412, 2695
Haguenauer P., Giuesalaga A., Butterley T., 2020, Proc. SPIE Conf. Ser. Vol. 11448, Adaptive Optics Systems VII. SPIE, Bellingham, p. 114481K
Kornilov V., Tokovinin A., Vozyacova O., Zaitsev A., Shatsky N., Potanin S., Sarazin M., 2003, Proc. SPIE, 4839, 837
Kornilov V. et al., 2014, PASP, 126, 482
Lac C. et al., 2018, Annales Geosci. Model Dev., 11, 1929
Lafore J.-P. et al., 1998, Annales Geophys., 16, 90
Lascaux F., Masciadri E., Hagelin S., 2009, MNRAS, 398, 1093
Lascaux F., Masciadri E., Hagelin S., 2010, MNRAS, 403, 1714
Lascaux F., Masciadri E., Hagelin S., 2011, MNRAS, 411, 693
Lascaux F., Masciadri E., Fini L., 2015, MNRAS, 449, 1664
Lyman R., Cherubini T., Businger S., 2020, MNRAS, 496, 4734
Madurovicz A., Macintosh B., Ruffio J.-B., Chilcote J., Bailey V., Poyneer L., Nielsen E., Norton P., 2018, Proc. SPIE Conf. Ser. Vol. 10703, Adaptive Optics Systems VI. SPIE, Bellingham, p. 107036E
Masciadri E.& Egner S., 2006, PASP, 118, 1604
Masciadri E.& Jabouille P., 2001, A&A, 376, 727
Masciadri E., Vernin J., Bougeault P., 1999a, A&AS, 137, 203
Masciadri E., Vernin J., Bougeault P., 1999b, A&AS, 137, 185
Masciadri E., Vernin J., Bougeault P., 2001, A&A, 365, 699
Masciadri E., Avila R., Sanchez L. J., 2002, A&A, 382, 387
Masciadri E., Feldt M., Hippler S., 2004a, ApJ, 613, 572
Masciadri E., Avila R., Sanchez L. J., 2004b, Rev. Mex. Astron. Astrofis., 40, 3
Masciadri E., Lascaux F., Fini L., 2013, MNRAS, 436, 1968
Masciadri E., Lascaux F., Fini L., 2017, MNRAS, 466, 520
Masciadri E., Turchi A., Martelloni G., 2019, preprint (arXiv:1911.02819)
Masciadri E., Martelloni G., Turchi A., 2020, MNRAS, 492, 140
Milli J. et al., 2019, preprint (arXiv:1910.13767)
Neichel B. et al., 2014, MNRAS, 440, 1002
Noilhan J., Planton S., 1989, Mon. Weather Rev., 117, 536
Ochs G. R., Wang T., Lawrence R. S., Clifford S. F., 1976, Appl. Optics, 15, 2504
Osborn J., Sarazin M., 2018, MNRAS, 480, 1278
Osborn J. et al., 2018, MNRAS, 478, 825
Pedichini F. et al., 2016, Proc. SPIE Conf. Ser. Vol. 9908, Ground-based and Airborne Instrumentation for Astronomy VI. SPIE, Bellingham, p. 990832
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Rabien S. et al., 2019, A&A, 621, A4
Rigaut F. et al., 2014, MNRAS, 437, 2361
Rigaut F. et al., 2020, Proc. SPIE Conf. Ser. Vol. 11447, Adaptive Optics Systems VIII. SPIE, Bellingham, p. 114471R
Ruggiero F. H., De Benedictis D. A., Lefevre R. J., Early S. A., 2002, 20th Conference on Weather analysis and Forecasting, DoD High Performance Modernization Program Users Group Conference. p. 10
Sarazin M., Roddier F., 1990, A&A, 227, 294
Sheperd H. W., Osborn J., Wilson R., Butterley T., Avila R., Dhillon V. S., Morris T. J., 2014, MNRAS, 437, 3568
Stein J., Richard E., Lafore J. P., Pinty J. P., Asencio N., Cosma S., 2000, Meteorol. Atmos. Phys., 72, 203
Trinquet V. J., Vernin J., 2007, Environ. Fluid Mech., 7, 397
Turchi A., Martelloni G., Masciadri E., 2018, Proc. SPIE Conf. Ser. Vol. 10703, Adaptive Optics Systems VI. SPIE, Bellingham, p. 107036H
Turchi A., Masciadri E., Pathak P., Kasper M., 2020, MNRAS, 497, 4910
VanZandt T. E., Green J. L., Gage K. S., Clark W. L., 1978, Radio Sci., 13, 819
Wu S. et al., 2020, PASP, 132, 084501
Ye Q. Z., 2011, PASP, 123, 113

APPENDIX A: ON T_X AND FTS

We report here a few notes related to a few T_X to avoid misunderstandings. In Section 1 it has been said that the computation time required to perform the forecast has to be $\leq (T_{disp} - T_{start})$. This does not means that the end of the computation time T_{comp} necessarily matches with T_{disp}. The computation time is related to the hardware characteristics and the model configuration. In our operational systems, for example, at present we can run a forecast with $\Delta T_{comp} = (T_{comp} - T_{start})$ of the order of 3 hours to forecast the coming night for a sub-kilometric horizontal resolution (500 m) and a model configuration as described in Section 3 with relatively light hardware (not superclusters). T_{disp} is the time in which the Observatory requires that forecasts are delivered and can be different for different observatories. As a first approximation, we have an interest in moving towards short ΔT_{comp}. However, it is not said that system A is preferable to system B just because ΔT_{comp} is shorter, as claimed by some authors (Osborn & Sarazin 2018). This has been explained extensively and illustrated with examples in Masciadri et al. (2019).

Also, the fact that the shorter $(T_{ini} - T_{fc})$ is, the better the model performance is only a qualitative statement. Other factors play a role in the forecast process, such as the quality of the initialization data, which is not the same at all synoptic hours (00:00, 06:00, 12:00, 18:00) UT. Initialization data calculated at 06:00 UT and 18:00 UT are worse than those calculated at 00:00 UT and 12:00 UT. Therefore, in Fig. 1 for example, a shorter $(T_{ini} - T_{fc})$ does not necessarily lead to better model performance. At the same time, given two sites A and B, with $(T_{ini} - T_{fc})_A > (T_{ini} - T_{fc})_B$, it does not necessarily follow that model performance above B is better than that above A, as the local characteristics play a role in the forecast process.

APPENDIX B: DEFINITIONS OF CONTINGENCY TABLES, PC, POD, AND EBD

Table B1 is an example of a generic 3 × 3 contingency table where the observations and simulations are divided into three categories delimited by two thresholds. PC, POD_i, and EBD can be defined using $a, b, c, d, e, f, g, h, i$ (number of times at which an observation and a simulation fall inside each category) and N (total events). The percentage of correct detection PC is defined in equation (B1), where PC = 100 per cent is the best score; the probability of detecting the value of a parameter inside a specific range of values (POD_i) is given by equations (B2)–(B4), where POD_i = 100 per cent is the best score. The extremely bad detection (EBD) probability is given by equation (B5), where EBD = 0 per cent is the best score. For a total random prediction and in the case of a 3 × 3 contingency table we have $a = b = ... = i = N/9$ and PC = POD_i = 33 per cent and EBD = 22.2 per cent.

$$PC = \frac{a+e+i}{N} \times 100; \quad 0 \text{ per cent} \leq PC \leq 100 \text{ per cent}, \tag{B1}$$

$$POD(\text{event}_1) = \frac{a}{a+d+g} \times 100; \quad 0 \text{ per cent} \leq POD \leq 100 \text{ per cent}, \tag{B2}$$

$$POD(\text{event}_2) = \frac{e}{b+e+h} \times 100; \quad 0 \text{ per cent} \leq POD \leq 100 \text{ per cent}, \tag{B3}$$

$$POD(\text{event}_3) = \frac{i}{c+f+i} \times 100; \quad 0 \text{ per cent} \leq POD \leq 100 \text{ per cent}, \tag{B4}$$

$$EBD = \frac{c+g}{N} \times 100; \quad 0 \text{ per cent} \leq EBD \leq 100 \text{ per cent}. \tag{B5}$$

Table B1. Generic 3 × 3 contingency table.

Intervals	OBSERVATIONS 1	2	3	Total
MODEL				
1	a (hit 1)	b	c	$a+b+c$ 1 (Model)
2	d	e (hit 2)	f	$d+e+f$ 2 (Model)
3	g	h	i (hit 3)	$g+h+i$ 3 (Model)
Total	$a+d+g$ 1 (OBS)	$b+e+h$ 2 (OBS)	$c+f+i$ 3 (OBS)	$N=a+b+c+d+e+f+g+h+i$ Total of events

APPENDIX C: OROGRAPHY

Table C1 and Fig. C1 report the extension, resolution, and total surface of the three domains forming the model orography in grid-nesting configuration (see Section 3).

Table C1. Meso-NH model grid-nesting configuration. The second column shows the number of horizontal grid points, the third column the domain extension, and the fourth column the horizontal resolution ΔX.

Domain	Grid points	Domain size (km)	ΔX (km)
Domain 1	80 × 80	800 × 800	$\Delta X = 10$
Domain 2	64 × 64	160 × 160	$\Delta X = 2.5$
Domain 3	120 × 100	60 × 50	$\Delta X = 0.5$

APPENDIX D: MEASURED AND FORECAST PARAMETERS: DENSITY FUNCTION MAPS

Fig. D1 shows the density function maps associated with the scattering plots reported in Fig. 3. We highlight the fact that the role of the density function map is only to put in evidence the peak of the distribution to retrieve potential bias effects that are not clearly visible in Fig. 3. Also it helps to put in evidence the density of values that are more or less correlated. Such a function is built with an interpolation of the colour levels assuming an arbitrary minimum threshold and a defined frequency of isolines. Such a figure of merit is therefore less efficient than scattering plots (see Fig. 3) to provide evidence for the effective dispersion of data (i.e. the dispersion of the individual pairs of points). Indeed, by selecting a suitable minimum threshold, the dispersion of values might appear much thinner than it is in reality. This is the reason why we prefer to report both figures (Figs 3 and D1).

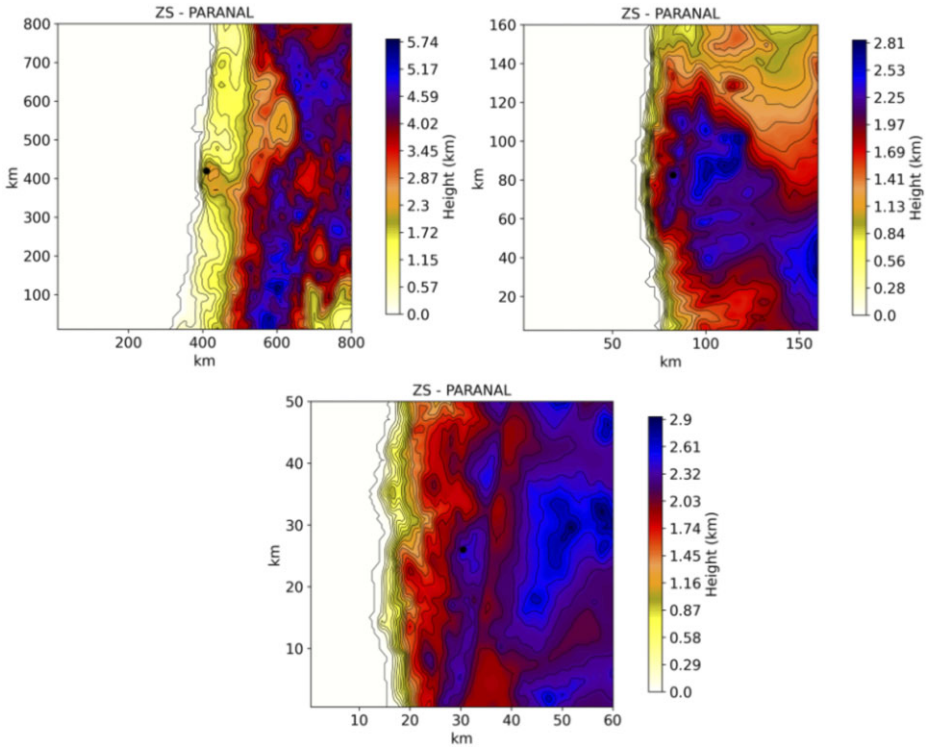

Figure C1. Orography of the Cerro Paranal summit. Top left: domain 1; top right: domain 2; bottom: domain 3. The black dot in each panel indicates the Cerro Paranal position. See Table C1.

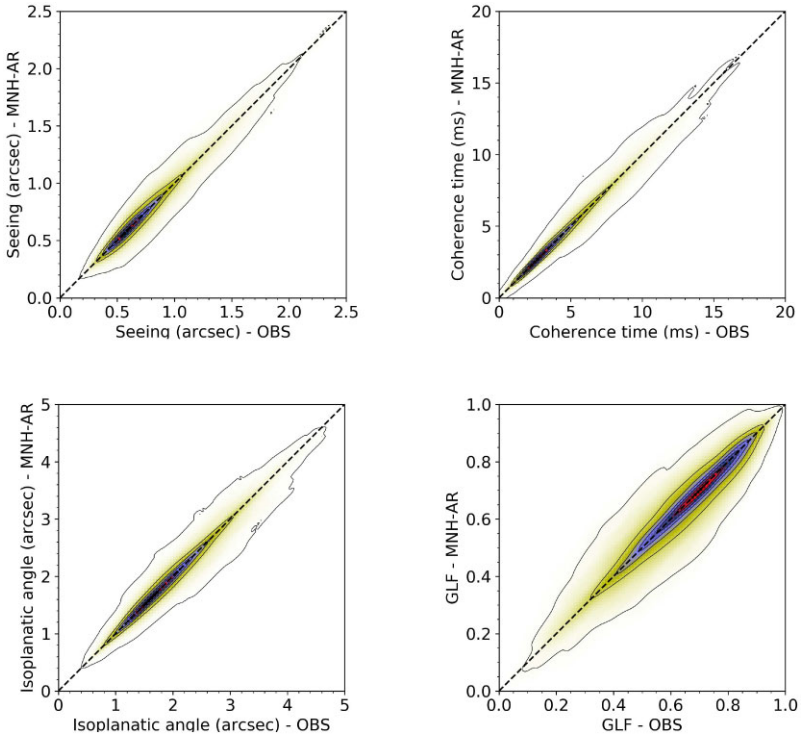

Figure D1. Density function map related to Fig. 3.

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