Seeing values for LSST strategy simulations

The opsim4 operations simulation program for the LSST astronomical survey uses a database of seeing values covering the range of times to be simulated. I describe the creation of such a database using Dual Image Motion Monitor (DIMM) data collected at Cerro Pachon from 2014-03-17 to 2018-03-21. This data has significant gaps, so I model the data and generate artificial data in the gaps according to the model. The model consists of a sinusoidal variation with a period of one year, an autoregressive (AR1) model for variations in mean seeing from one night to the next, and another AR1 model for variations on a 5 minute timescale. I create two databases according to this procedure, one based on DIMM data from 2008-01-01 to 2018-12-31, and the other for 2005-01-01 to 2015-12-31, and run opsim4 simulations using each. I also run an otherwise identical simulation using the default seeing database. The results show a variation in mean point spread function (PSF) width with right ascension with peak-to-peak differences of 17% and 16%, respectively, compared to a 2% variation in the default seeing database.

1. Introduction

The Large Synoptic Survey Telescope (LSST) is a telescope currently under construction on Cerro Pachon, in Chile. It will spend 10 years performing an astronomical survey, taking repeated images across a large fraction of the sky visible from Cerro Pachon. It will take images of any given area of the sky multiple times, spread across many nights, so that transient objects may be detected and characterized.

Turbulence in the Earth's atmosphere causes short time-scale variations in the index of refraction of the air. These variations place limits on the sharpness of astronomical images taken by telescopes on the surface of the Earth; this limit is called the "seeing", typically measured as the angular full width at half maximum (FWHM) of the image of a point source, the "point spread function" (PSF), that would be taken by an ideal instrument. The seeing is a property of the weather, and as such is correlated with the location, time of year, and transient weather patterns. Els et al. (2009), for example, measure a significant variation in seeing with time of year at Cerro Tololo, a site $\sim$ 10 km from Cerro Pachon.

The `opsim4` operations simulation computer program generates a database of exposures plausible for an execution of the survey. Each exposure in the database includes several parameters, including the time the exposure was taken, the depth of the image (the brightness of the faintest objects detected at a given signal to noise ratio), and the delivered PSF FWHM. The LSST project and science groups use these databases of simulated exposures to evaluate the data quality that would result from different operations strategies. Such evaluations can then be used both to select among candidate observing strategies, and set expectations for the scientific usefulness of the LSST data set.

To calculate the depth and PSF FWHM of each simulated exposure, the simulator must have a value for the atmospheric seeing at the time the image was taken. It takes these values from a simple database table, which provides atmospheric seeing values at a set of times.

The seeing data is not only used to calculate the data quality values for each exposure in the final database; it may also affect the scheduling of the exposures themselves. Some science programs are more sensitive to the seeing than others, and the scheduler can use this data in its choice of which exposure to take at a given time. For example, a less urgent exposure needed for one program may be taken in preference to a more urgent exposure needed for another, if the seeing is too poor for the data to be useful for the more urgent exposure, but adequate for the less urgent one.

Therefore, the details of the seeing database used by `opsim4` can therefore affect the results in several ways:

- The global quality of the survey is strongly affected by the contents of the seeing database. If the average seeing in the seeing database is worse, the average delivered PSF FWHM in the images that comprise the survey will be worse, as will the depth of the survey.
- The accessible area in the sky varies with a period of one year, which corresponds to the yearly seasonal variation in the seeing. For example, the same area on the sky

can be imaged in January every year, and a different area every July. If the seeing is better in January than in July of every year, then the data quality in the area of sky accessible in January will be better than that accessible in July.

- The autocorrelation of seeing over time will also affect the data quality of light curves of transient objects: if the seeing is weakly correlated over timescale similar to the duration of a transient event, then the quality of different points on the light curve will be uncorrelated. On the other hand, if the autocorrelation of the seeing over time is strong on the timescale of the event, then it is more likely for the seeing to be either good or poor over the whole duration of the event.
- The autocorrelation of the seeing over time on timescales similar to the time between one exposure and the next will affect the ability of the scheduler to react appropriately to changes in seeing.

The seeing conditions on Cerro Pachon have been monitored since 2004 using a Dual Image Motion Monitor, or DIMM. A DIMM measures the position of a star through two neighboring paths through the atmosphere, typically separated by $\sim$ 10 cm. The difference in positions between these two paths indicates the variability in measured position due to turbulence on that spatial scale. The Fried parameter, the diameter of a circular aperture over which the RMS wavefront error induced by atmospheric turbulence is one radian, can be derived directly from DIMM measurements [Fried (1965); Martin (1987); Tokovinin (2002)].

The archive of DIMM data for Cerro Pachon records seeing values derived for 0.5μm light using a Kolmogorov turbulence model, which may be pessimistic. Tokovinin (2002) provides a formula for approximating a more realistic von Kármán model, provided one can estimate the outer scale of the turbulence.

The default seeing database used by `opsim4` version 081217 was artificially generated from a model derived from a more limited set of data from the Cerro Pachon DIMM, and repeats with a period of two years.

Observing strategy simulation for the Dark Energy Survey (DES) [Dark Energy Survey Collaboration et al. (2016)] had a similar requirement. `obstac` [Neilsen & Annis (2014)], the DES operations scheduler and simulator, used seeing data sets generated using a model derived from data from the DIMMs on Cerro Tololo [Neilsen (2012)]. The model used by `obstac` included both a seasonal component and a short timescale autoregressive model, producing seeing values on 5 minute intervals.

Plots in this note were generated using `jupyter` notebooks that can be found in github: `https://github.com/LSSTDESC/obs_strat/tree/master/doc/seeing`. The `simsee` application itself can be found in the same product: `https://github.com/LSSTDESC/obs_strat/tree/master/code/simsee`. Data files not included in the `github` product can be found in the `/global/project/projectdirs/lsst/survey_sims/input/seeing` directory on `cori.nersc.gov`.

2. Methods

2.1. Overview

The following procedure was followed in generating new seeing databases and exploring their effects on `opsim4` simulations:

- Obtain the Cerro Pachon DIMM data, and explore it interactively, as provided. See section 2.2.
- For each DIMM measurement, calculate the Fried parameter, r_0, and seeing based on the von Kármán model using the correction given in Tokovinin (2002) and an outer scale of $\mathcal{L}_0 = 30$ meters, based on the measurement reported in Ziad et al. (2000).
- Resample the DIMM data to obtain a data set sampled on 5 minute intervals.
- Interactively explore long time-scale variability, and fit a sine with a period of 1 year to the nightly mean value of $\log(r_0)$.
- Interactively explore the nightly residuals of $\log(r_0)$ (after subtraction of the seasonal model), and fit the residuals using autoregressive (AR1) models on each uninterrupted sequence of consecutive nights with DIMM data. Derive a global AR1 model for nightly residuals using a weighted average of the parameters as derived from each sequence of nights.

- Interactively explore the short time-scale residuals of $\log(r_0)$ (after subtraction of the nightly mean values), and fit the residuals using a second autoregressive (AR1) model on each consecutive sequence of values in the resampled data. (Gaps in DIMM data during the night result in breaks between sequences in the resampled data.) Derive a global AR1 model for short time-scale residuals using a weighted average of the parameters as derived from each sequence.
- Generate two seeing databases, using two different subsets of the Cerro Pachon DIMM data. Use the derived models to fill in gaps in the DIMM data with values randomly generated according to the derived seasonal, nightly, and short-timescale AR models.
- Run three `opsim4` simulations: one using the default seeing database, and one for each of the newly generates seeing databases.
- Compare the results of these three simulations.

2.2. Cerro Pachon DIMM data

Bustos (2018) kindly provided an the Cerro Pachon DIMM data in the form of two text tables, each with a timestamp and an airmass-corrected seeing value for a wavelength of 0.5 μm, derived using a Kolmogorov seeing model.

Figure 1 shows the variation in reported FWHM seeing values with time. The regular extremes of good and poor seeing, shortly after the start and midpoint of each year, match well with anecdotal experience, and indicate a significant seasonal component. There are also noticeable long-term trends, but on a timescale comparable to or greater than the range of the data, so no attempt is made to model these longer term trends here. This feature is also evident in the autocorrelation function of the nightly means,

The Fried parameter, r_0, was then calculated for each reported value by inverting equation 5 of Tokovinin (2002). Figure 2 shows the distribution of measured values for the r_0, in meters (left) and $\log(r_0)$ (right), together with best fit normal distributions. Neither distribution is precisely normal, but $\log(r_0)$ is noticeably closer to normal.

Figure 3 shows the time series of DIMM measurements for three nights, chosen randomly from among nights with good DIMM coverage.

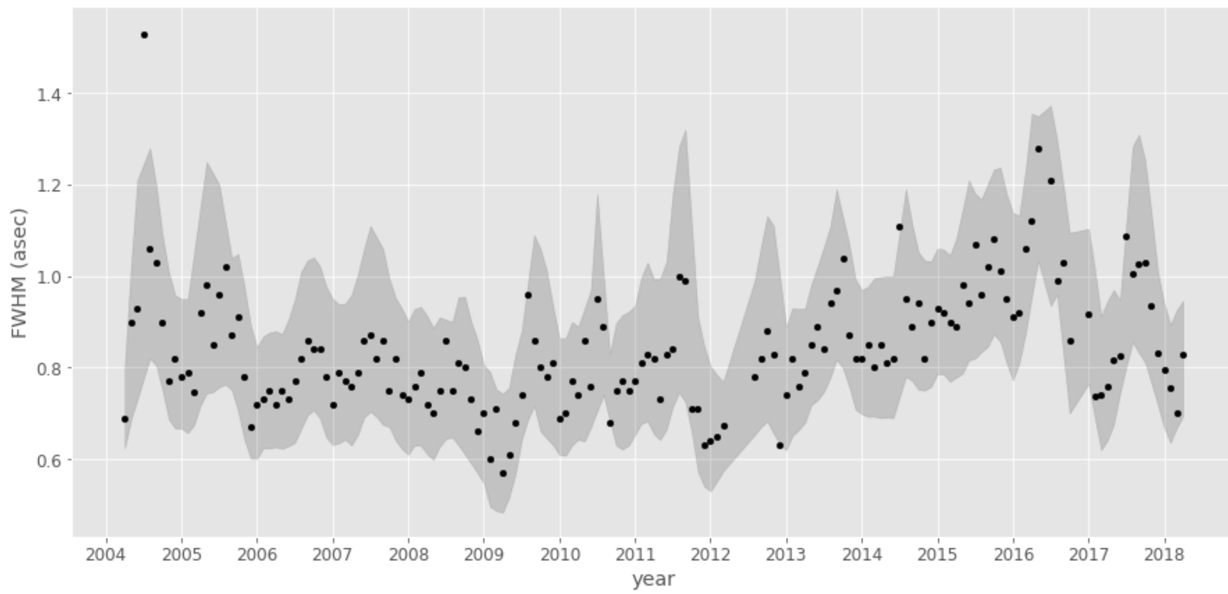

Figure 1. Points show the mean FWHM reported by the DIMM for each month. The upper and lower bounds of the dark gray band are LOWESS smoothed curves derived from the 1st and 3rd quartiles of this data.

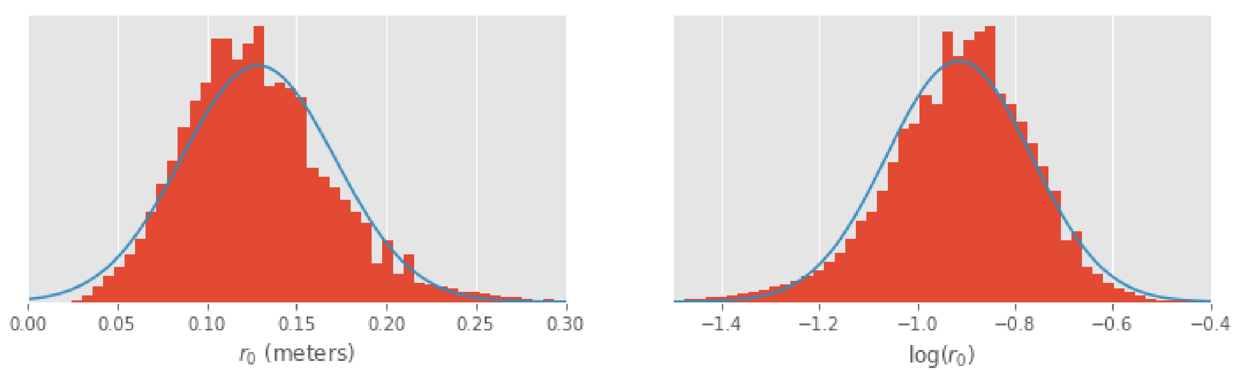

Figure 2. The distribution of the Fried parameter, r_0 in meters (left), and $\log(r_0)$ (right), together with best fit normal distributions.

2.3. Seasonal fit

Figure 3. The time series of DIMM measurements for three nights, chosen randomly from among nights with good DIMM coverage.

a	-0.9170	corresponds to a seeing of 0.83″ (Kolmogorov)
c	0.048	corresponds to a peak to peak amplitude of 0.18″
d	18.8	best seeing on January 18, worst on July 21
Nightly L1	0.2	
Nightly e	0.084	
5 minute L1	0.7	
5 minute e	0.052	

Table 1. Parameters for the least squares fit of Pachon DIMM data to equation 1, the nightly autoregressive model, and the short time-scale autoregressive model.

The first stage in creating a model for the seeing data was to fit a sine curve (plus a constant) with a period of one year to $\log(r_0)$ (equation 1).

$$\log(r_0) = a + c \times \cos\left((\text{day} - d) \times \frac{2\pi}{365.24217}\right) \tag{1}$$

A sine was chosen as the simplest periodic function. Table 1 shows the best fit values. Figure 4 shows the distribution of the mean $\log(r_0)$ values for each month, before and after subtraction of the seasonal model. Before subtraction of the model, months near the middle of the year have obviously worse seeing, an effect not visible after subtraction. Figure 5 shows the autocorrelation functions of the monthly mean $\log(r_0)$ values. The seasonal effect is again prominent before subtraction of the seasonal fit. Longer timescale variations are still apparent after subtraction of the seasonal model, but there is no obvious periodic structure.

2.4. Nightly variation in seeing

After subtraction of the seasonal variation in $\log(r_0)$, significant night to night correlation remains. The modeled here as a first-order autoregressive process [Cryer & Chan (2008)], also referred to as an AR(1) process or a damped random walk [Kelly et al. (2009)], described in equation 2. Here, y_t is the difference between the mean $\log(r_0)$ on that night and the seasonal model for $\log(r_0)$. L1 is the regressive term, the model parameter that

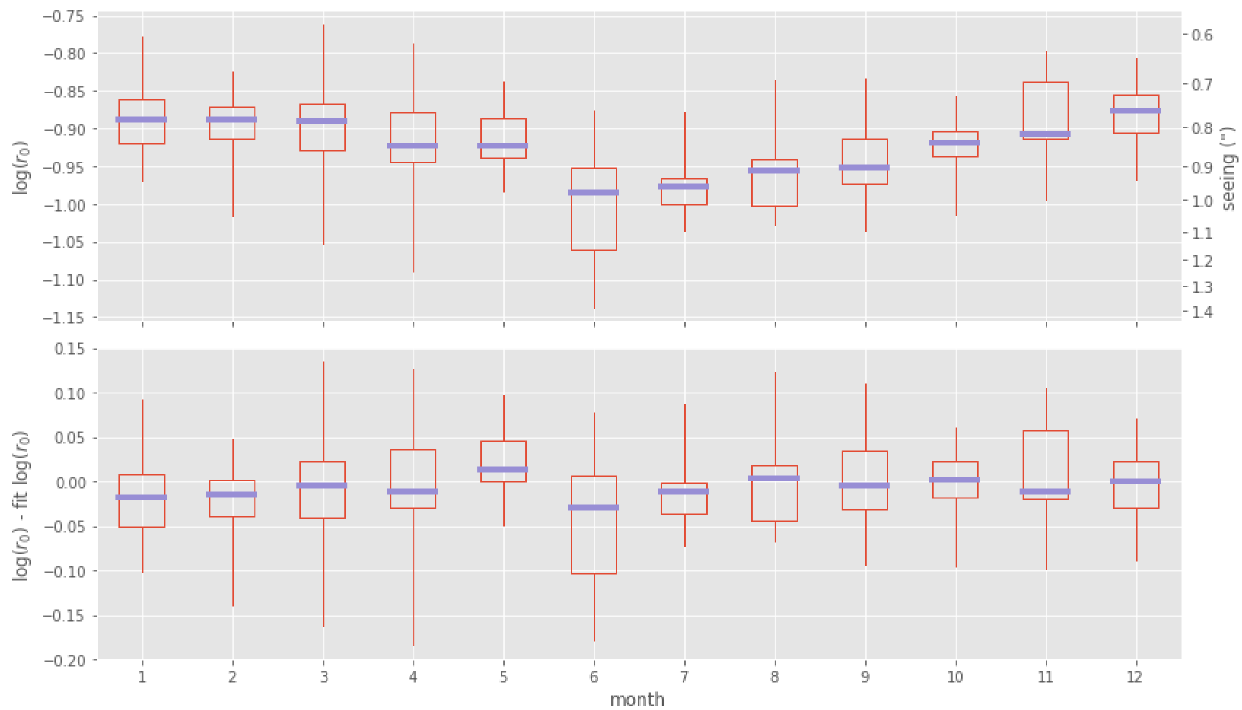

Figure 4. The top plot shows the distributions of mean monthly values for each month. The blue bar shows the median mean value for that month, the box the 1st and 3rd quartiles, and the whiskers the full range. The bottom plot shows the distributions after subtraction of the seasonal fit to $\log(r_0)$.

represents the correlation between one night and the next, and e_t is the "innovation", analogous to the step sizes in a random walk.

$$y_t = \mathrm{L1} \times y_{t-1} + e_t \tag{2}$$

There are many nights in the Cerro Pachon DIMM data set with no data. The tool used to fit the AR1 model, from the python `statsmodels` module [Seabold & Perktold (2010)], does not handle missing data. To fit the AR1 model, I divided the full data set into sequences of consecutive nights without missing data, performed separate fits on each sequence, and accepted the mean values provided by the models, weighted according to the reported uncertainty in each model fit. Table 1 lists the resultant fit parameters.

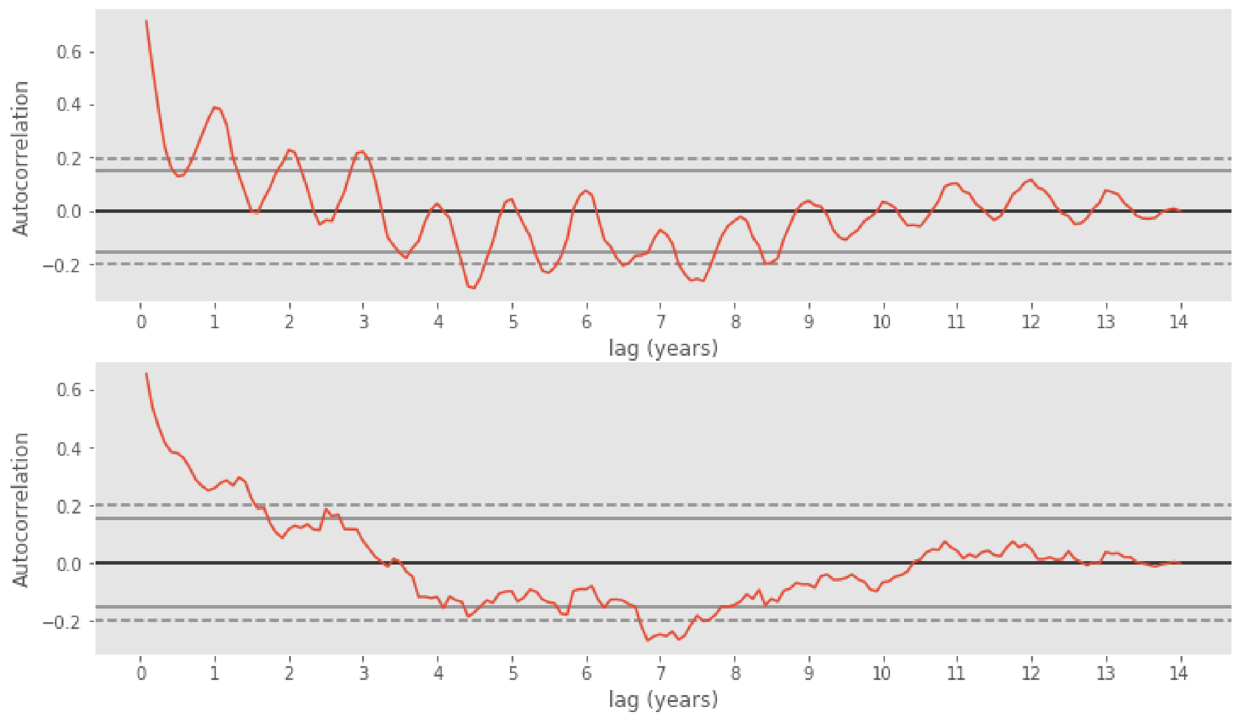

Figure 5. The upper and lower plots show the autocorrelation function of the mean $\log(r_0)$ by month, before and after subtraction of the seasonal model, respectively. The solid and dashed gray lines show the 95% and 99% ranges for uncorrelated data.

The distribution values in a sequence of points generated by an AR1 process is a normal distribution with a variance given by equation 3[Cryer & Chan (2008) eqn. 4.3.3].

$$\sigma_y^2 = \frac{\sigma_e^2}{1 - \mathrm{L1}^2} \tag{3}$$

Figure 6 shows the expected distribution given the fit model parameters over-plotted over the actual histogram of $\log(r_0)$ - seasonal fit $\log(r_0)$. The distribution expected from the AR1 fit is sharper than the measured one, and does not capture the tail on the low-$\log(r_0)$ side of the distribution. The later is a fundamental limitation of the model. The sharper distribution likely arises from the fit of a collection of sub-sequences of nights, rather than a full, uninterrupted data set: figure 1 clearly shows variations on timescales of months to years, too long to be captured by the sub-sequences of nights to which the AR1 model was fit.

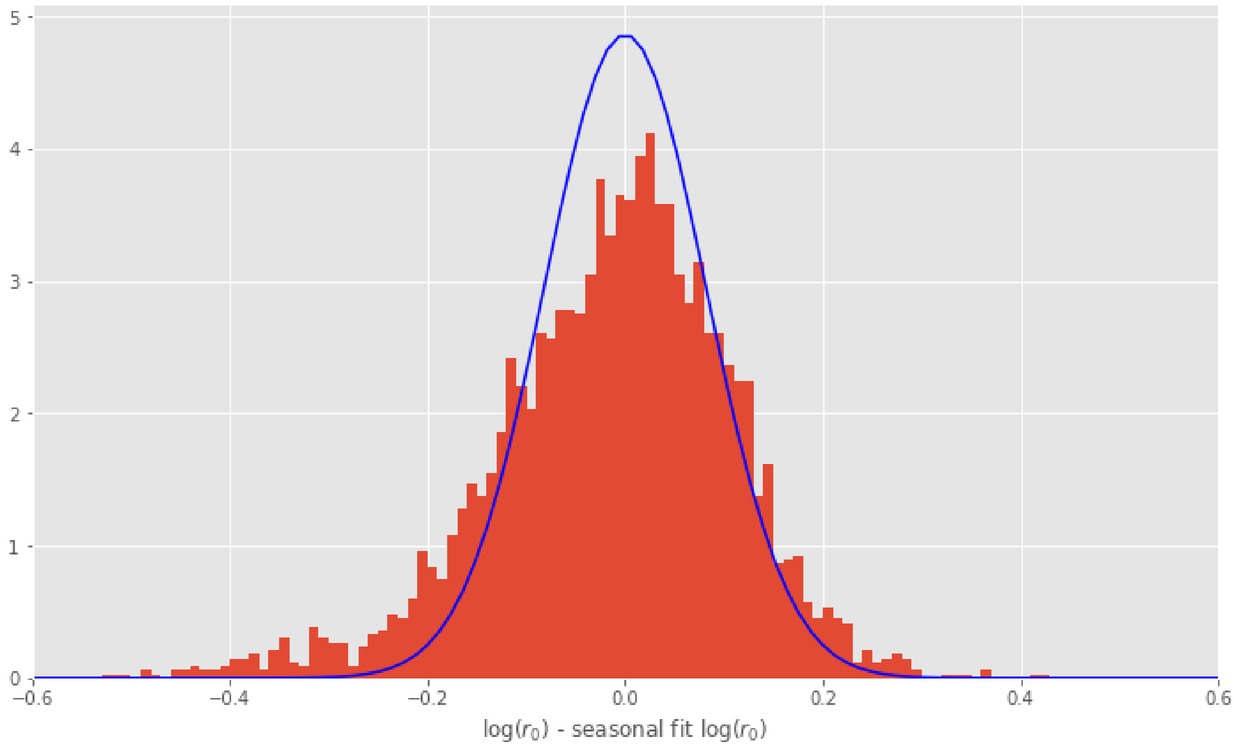

Figure 6. The histogram of differences between nightly mean $\log(r_0)$ and the seasonal fit, overplotted by the result that would be expected by the fit AR1 model.

2.5. Short timescale variation in seeing

In addition to varying on a nightly basis, seeing varies on much shorter timescales. The short-timescale variations are modeled using an AR1 model as well. The raw DIMM data is sampled irregularly, slightly more frequently than once every 5 minutes. I therefore resample the points onto exact 5 minute intervals, and divide it in into sub-sequences of consecutive uninterrupted exposures, similar to the procedure for nightly data. Table 1 lists the resultant fit parameters.

2.6. Correction from Kolmogorov to von Kármán turbulence

The raw data provided by the Cerro Pachon DIMM archive provides seeing data calculated using a Kolmogorov model for the turbulence in the atmosphere. This data was used to work backward to the Fried parameter, r_0, which was then modeled. To obtain simulation seeing values from the r_0 model, I use the approximation given in equation 4, provided by Tokovinin (2002).

$$\left(\frac{\mathrm{FWHM}_{vK}}{\mathrm{FWHM}_K}\right)^2 \approx 1 - 2.183\left(\frac{r_0}{\mathcal{L}_0}\right)^{0.356} \tag{4}$$

I use a value of $\mathcal{L}_0 = 30$ meters, based on the value of $28.4^{+25.0}_{-13.3}$ meters reported by Ziad et al. (2000) for Cerro Pachon. This corresponds to a 22% improvement in seeing when converting from a Kolmogorov to a von Kármán turbulence model and a typical value of r_0, but the range given is from 18% to 30%. Furthermore, the value reported was measured data from only a few nights of data, and is likely to be strongly dependent on weather.

2.7. Seeing data generation

Much of the data to be used by an `opsim4` simulation can be copied directly from the historical DIMM data, after conversion from a Kolmogorov to a von Kármán turbulence model and application of an offset in time by an integer number of years. The gaps can then be filled in using the model.

For sequences of nights with no data, mean values for each night are calculated using the seasonal model (equation 1) and the nightly AR1 model (equation 2) with the last night of data with DIMM data and randomly generated values of e_t. For sequences of short time-scale (5 minute interval) points, artificial data is generated similarly, using the nightly mean, the last good DIMM data point, and random values of e_t.

Two different seeing databases were generated: one using a 14 year offset (such that the 2022-01-01T00:00:00Z data point in the data set is copied from the 2008-01-01T00:00:00Z DIMM data), and one using a 17 year offset, such that we take full advantage of all available DIMM data. Note that there is overlap between the DIMM data used by these two database, so the results are not uncorrelated.

2.8. `opsim4` simulations

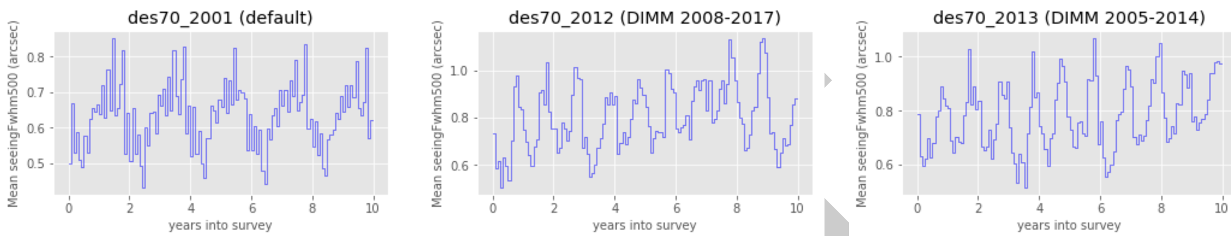

Figure 7. Each plot shows the variation in seeing with time produced by a different run of `opsim4`. The leftmost plot shows the seeing for the default seeing database, and the remaining two show the seeing for each of the seeing databases produced by `simsee`.

Three separate simulations were run using `opsim4-081217`, as distributed from https://hub.docker.com/r/oboberg/opsim4/. Each simulation was run for a full 10-year LSST survey, with the default configuration except for the seeing database.

`des70_2001`: a 10-year simulation using the seeing database provided in the `opsim4-081217` container, used as a reference.

`des70_2012`: a simulation using the `simsee_pachon6` seeing database, which uses DIMM data from 2008-01-01 to 2018-12-31 to to simulate 2022-01-01 to 2032-12-31. This simulation is otherwise identical to `des70_2001`.

`des70_2013`: a simulation using the `simsee_pachon7` seeing database, which uses DIMM data from 2005-01-01 to 2015-12-31 to to simulate 2022-01-01 to 2032-12-31. This simulation is otherwise identical to `des70_2001`.

3. Results

Figure 7 shows the seeing as a function of time, as recorded in the databases produced by each of the three runs of `opsim4`. The two year periodicity of the seeing that results from the default two year input database is apparent in the leftmost plot in the figure. Yearly periodicity, expected from the seasonal variation in the DIMM data and model, is apparent in the plots from the other two runs. The average seeing in the default model is also better than either of the revised simulations: the mean FWHM in the revised simulations are 13% and 14% wider than the default.

Figure 8 shows maps of the mean seeing in the LSST wide-fast-deep (WFD) survey from each simulation. Degradation is apparent near the northern and southern edges of all

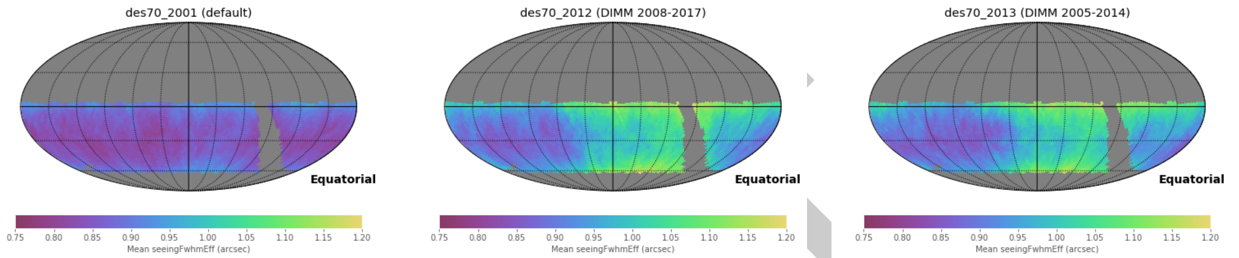

Figure 8. Each plot shows the map of mean seeing produced by a different run of `opsim4`. The leftmost plot shows the seeing for the default seeing database, and the remaining two show the seeing for each of the seeing databases produced by `simsee`.

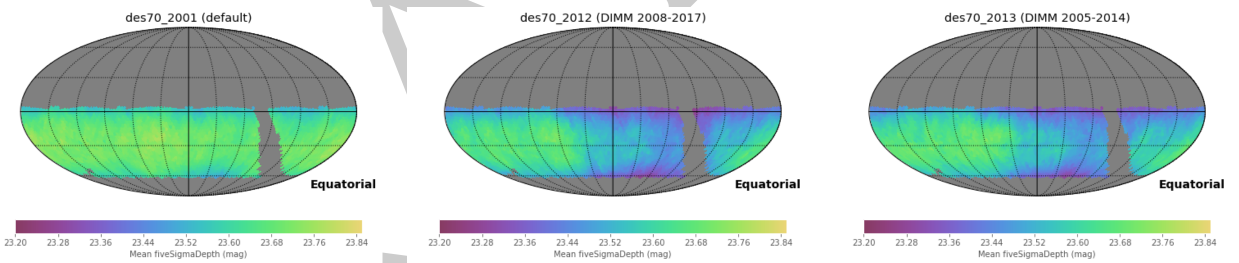

Figure 9. Each plot shows the map of mean depth produced by a different run of `opsim4`. The leftmost plot shows the depth for the default seeing database, and the remaining two show the depth for each of the seeing databases produced by `simsee`.

three simulations. This is expected, because these areas are never at low airmass from Cerro Pachon. The seeing in the default simulation shows no obvious trends otherwise. In the simulations that use the revised seeing databases, however, there is an obvious large scale trend with R.A. This same trend is also apparent in the plot of the mean seeing as a function of R.A. in the left plot of figure 10: seeing in fields that transit when the seeing is poor ($180^\circ < \text{R.A.} < 330^\circ$) have a mean FWHM 17% wider than those that transit when the seeing is good ($60^\circ < \text{R.A.} < 150^\circ$).

The variation in seeing corresponds to a variation in depth, shown in figure 9 and the right-hand plot in figure 10: the $5-\sigma$ magnitude limit in fields that transit when the seeing is poor ($180^\circ < \text{R.A.} < 330^\circ$) is ~ 0.12 shallower than those that transit when the seeing is good ($60^\circ < \text{R.A.} < 150^\circ$).

4. Discussion

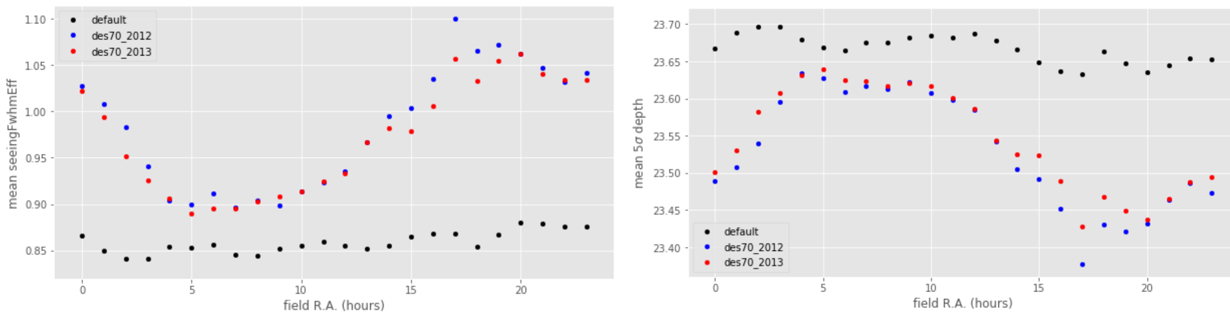

Figure 10. The seeing and depth as a function of RA for different runs of `opsim4`.

The use of real seeing data (and a more elaborate model for times when such data is not available) in operations simulations demonstrates a significant, large angular scale variation in seeing (and therefore depth) using the current strategy. The impact of this variation on science results needs to be carefully evaluated and, if warranted by the science, adjustments to the strategy made to mitigate these effects.

Such mitigation strategies will necessarily come at a cost. The current strategy is designed to observe fields when they are near transit. Such a strategy optimizes the quality of data taken at any given time: a field observed near transit at a given declination will have a better FWHM than another at the same declination, but further from transit. Observing fields dear transit, however, necessarily maps time of observation directly to sidereal time, which is correlated with time of year, and therefore the seasonal variation in seeing.

This correlation can be reduced by observing fields when they are further from transit, depending on seeing conditions. For example, the strategy could be modified to select fields in the (180° < R.A. < 330°) range whenever such fields are at an airmass better than some threshold and the seeing is better than average, and in the (60° < R.A. < 150°) range when it is worse than average. Such a strategy will tend to even out the extremes in the variation. However (if the strategy maintains the global distribution in declination) these exposures will be at higher airmasses than those that would have been taken at transit, which will degrade the overall mean image quality.

A compromise will need to be made. The effect on image quality is not linear with zenith distance (or time from transit), but is shallow very close to transit, and degrades more rapidly as the angle increases: a mild deviation from the transiting strategy may only have a mild effect on the mean image quality.

5. Future work

Although an improvement over the default seeing database, the revised seeing databases presented here leave significant room for improvement. Some refinements that should be explored include:

- Modeling the long term (multi-year) trends in the DIMM seeing. Currently the model regresses to the average seeing (for the specific time of year) quickly, rather than keeping the multi-year trend apparent in the data.
- Rigorous evaluation of higher-order ARMA models using a formal criteria (either Akaike's Information Criterion (AIC) or a Bayesian Information Criterion (BIC))[Cryer & Chan (2008) pp. 130-132], rather than the AR1 model used here. The AR1 model was selected due to its simplicity and apparent effectiveness after informal exploration; additional terms and/or a moving average component may be warranted.
- Modeling the short term, nightly, seasonal, and long term components as a single seasonal ARMA model, following the formalism described in Cryer & Chan (2008) chapter 10. Rather than fit each element separately (as has been done here), this approach incorporates long-term effects by including additional terms in the autoregressive equation.
- Modeling using a continuous ARMA model (CARMA) [Brockwell & Davis (1996) pp. 344-348] rather than the discrete ARMA model used here. Such models are significantly more complex and lack the well developed software tools, but naturally handle the irregularly sampled nature of the DIMM data.
- Creating a better model of the poor seeing tail in the distribution of seeing values, either as an additional component or by transforming the DIMM's $\log(r_0)$ distribution.
- Interactive exploration and informal experience suggest that the seeing has a systematic variation with the time of night, in particular that the seeing is slightly worse shortly after sunset. This needs to be studied further, and perhaps modeled as well.

In addition to modeling the seeing, improved modeling of the effect of clouds in survey data quality should also be studied.

Finally, observing strategy adjustments may need to be developed to address non-uniformity introduced by seasonal seeing variations.

6. Conclusion

The full archive of data from the Cerro Tololo DIMM shows strong seasonal variations, and larger mean values for the seeing, than are present in the default input database used by the LSST `opsim4` operations simulator. Inclusion of an updated seeing database is therefore important for using `opsim4` to evaluate both the overall survey quality and also large scale variation in seeing and depth across the survey footprint.

Acknowledgments

I am grateful to Edison Bustos and other staff of the National Optical Astronomical Observatory (NOAO) for supplying the data used in this study.

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

The DESC acknowledges ongoing support from the Institut National de Physique Nucléaire et de Physique des Particules in France; the Science & Technology Facilities Council in the United Kingdom; and the Department of Energy, the National Science Foundation, and the LSST Corporation in the United States. DESC uses resources of the IN2P3 Computing Center (CC-IN2P3–Lyon/Villeurbanne - France) funded by the Centre National de la Recherche Scientifique; the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231; STFC DiRAC HPC Facilities, funded by UK BIS National E-infrastructure capital grants; and the UK particle physics grid, supported by the GridPP Collaboration. This work was performed in part under DOE Contract DE-AC02-76SF00515.

References

Brockwell, P. J., & Davis, R. A. 1996, Introduction to Time Series and Forecasting (Springer Science & Business Media)

Bustos, E. 2018, personal communication

Cryer, J. D., & Chan, K.-S. 2008, Time Series Analysis: With Applications in R (Springer Science & Business Media)

Dark Energy Survey Collaboration, Abbott, T., Abdalla, F. B., et al. 2016, MNRAS, 460, 1270

Els, S. G., Schöck, M., Bustos, E., et al. 2009, PASP, 121, 922

Fried, D. L. 1965, Journal of the Optical Society of America (1917-1983), 55, 1427

Kelly, B. C., Bechtold, J., & Siemiginowska, A. 2009, ApJ, 698, 895

Martin, H. M. 1987, PASP, 99, 1360

Neilsen, E., & Annis, J. 2014, in Astronomical Society of the Pacific Conference Series, Vol. 485, Astronomical Data Analysis Software and Systems XXIII, ed. N. Manset & P. Forshay, 77

Neilsen, Jr., E. H. 2012, in Astronomical Society of the Pacific Conference Series, Vol. 461, Astronomical Data Analysis Software and Systems XXI, ed. P. Ballester, D. Egret, & N. P. F. Lorente, 201

Seabold, S., & Perktold, J. 2010, in 9th Python in Science Conference

Tokovinin, A. 2002, PASP, 114, 1156

Ziad, A., Conan, R., Tokovinin, A., Martin, F., & Borgnino, J. 2000, ApOpt, 39, 5415