

Uncertainty-Aware Learning for Improvements in Image Quality of the Canada-France-Hawaii Telescope

Sankalp Gilda, Stark C. Draper, Sébastien Fabbro, William Mahoney, Simon Prunet, Kanoa Withington, Matthew Wilson, Yuan-Sen Ting, Andrew Sheinis

▶ **To cite this version:**

Sankalp Gilda, Stark C. Draper, Sébastien Fabbro, William Mahoney, Simon Prunet, et al.. Uncertainty-Aware Learning for Improvements in Image Quality of the Canada-France-Hawaii Telescope. 2021. hal-03388188

HAL Id: hal-03388188

https://cnrs.hal.science/hal-03388188

Preprint submitted on 20 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MNRAS 000, 1–25 (2015) Preprint 2 July 2021 Compiled using MNRAS LATEX style file v3.0

Uncertainty-Aware Learning for Improvements in Image Quality of the Canada-France-Hawaii Telescope

Sankalp Gilda[1],★ Stark C. Draper[2], Sébastien Fabbro[3], William Mahoney[4], Simon Prunet[4,9], Kanoa Withington[4], Matthew Wilson[4], Yuan-Sen Ting[5,6,7,8], and Andrew Sheinis[4]

[1]*Department of Astronomy, University of Florida, Gainseville, FL 32611, United States*
[2]*Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada*
[3]*National Research Council Herzberg, 5071 West Saanich Road, Victoria, BC, Canada*
[4]*Canada-France-Hawaii-Telescope, Kamuela, HI, United States*
[5]*Institute for Advanced Study, Princeton, NJ 08540, United States*
[6]*Department of Astrophysical Sciences, Princeton University, Princeton, NJ 08540, USA*
[7]*Observatories of the Carnegie Institution of Washington, 813 Santa Barbara Street, Pasadena, CA 91101, USA*
[8]*Research School of Astronomy & Astrophysics, Australian National University, Cotter Rd., Weston, ACT 2611, Australia*
[9]*Université Côte d'Azur, Observatoire de la Côte d'Azur, CNRS, Laboratoire Lagrange, France*

arXiv:2107.00048v1 [astro-ph.IM] 30 Jun 2021

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We leverage state-of-the-art machine learning methods and a decade's worth of archival data from the Canada-France-Hawaii Telescope (CFHT) to predict observatory image quality (IQ) from environmental conditions and observatory operating parameters. Specifically, we develop accurate and interpretable models of the complex dependence between data features and observed IQ for CFHT's wide field camera, MegaCam. Our contributions are several-fold. First, we collect, collate and reprocess several disparate data sets gathered by CFHT scientists. Second, we predict probability distribution functions (PDFs) of IQ, and achieve a mean absolute error of $\sim 0.07''$ for the predicted medians. Third, we explore data-driven actuation of the 12 dome "vents", installed in 2013-14 to accelerate the flushing of hot air from the dome. We leverage epistemic and aleatoric uncertainties in conjunction with probabilistic generative modeling to identify candidate vent adjustments that are in-distribution (ID) and, for the optimal configuration for each ID sample, we predict the reduction in required observing time to achieve a fixed SNR. On average, the reduction is $\sim 15\%$. Finally, we rank sensor data features by Shapley values to identify the most predictive variables for each observation. Our long-term goal is to construct reliable and real-time models that can forecast optimal observatory operating parameters for optimization of IQ. Such forecasts can then be fed into scheduling protocols and predictive maintenance routines. We anticipate that such approaches will become standard in automating observatory operations and maintenance by the time CFHT's successor, the Maunakea Spectroscopic Explorer (MSE), is installed in the next decade.

Key words: methods: statistical – methods: analytical – methods: observational – telescopes – instrumentation: miscellaneous

1 INTRODUCTION

Situated at the summit of the 4,200m volcano of Maunakea on the island of Hawaii, the Canada-France-Hawaii Telescope is one of the world's most productive ground-based observatories (Crabtree 2019). The productivity of CFHT is due, in part, to the exquisite natural image quality (IQ) delivered at the observatory's location on Maunakea. Image quality is key metric of observatory operations and relates directly to realized signal-to-noise ratio (SNR) as well as to achievable spatial resolution. SNR and spatial resolution, in turn, dictate the information content of an image. They thereby provide a direct measure of the efficacy of scientific observation.

The difference between the theoretically achievable and measured IQ can be attributed to air turbulence in the optical path. There are two sources of turbulence. The first is atmospheric. At the summit of Maunakea atmospheric turbulence is minimal due to the

★ E-mail: s.gilda@ufl.edu

© 2015 The Authors

smooth laminar flow of the prevailing trade winds and the height of the summit; this is the reason CFHT and other world-class observatories are located on Maunakea. The second is turbulence induced by local thermal gradients between the observatory dome itself (and the structures within) and the surrounding air. There have been continual improvements in the CFHT facility since 1979, many aimed at reducing this source of turbulence. We particularly make note of the December 2012 installation of dome vents. After a protracted mechanical commissioning period that lasted about 18 months, the vents came online in July of 2014. By allowing the (generally) hotter air within the observatory to flush faster, the vents accelerate thermal equalization. A schematic of the dome and the vents is provided in Figure 1. A listing of the temperature sensors marked in Figure 1 is provided in Table 1. Even given these improvements, and as is the case with all major ground-based observatories, the IQ attained at CFHT rarely reaches what the site can theoretically deliver.[1]

Our project is motivated by our strong belief that the ability to model and predict IQ accurately in terms of the exogenous factors that affect IQ would prove enormously useful to observatory operations. Observing time at world-class facilities like CFHT is oversubscribed many-fold by science proposals. Specifically at CFHT, good seeing time, defined as time when IQ is smaller than the mode seeing of $0.70''$ in the r-band, is oversubscribed roughly three-fold. Further, observations frequently either fail to meet, or exceed, the IQ requirements of their respective science proposals (Milli et al. 2019). Through accurate predictions we can better match delivered IQ to scientific requirements. We thereby aim to unlock the full science potential of the observatory. If we can predict the impact on IQ of the parameters the observatory can control (pointing, vent and wind-screen settings, cooling systems), then by adjusting these parameters and (perhaps) the order of imaging, we create an opportunity to accelerate scientific productivity. In this work we lay the groundwork for these types of improvements.

In this paper, we leverage almost a decade-worth of sensor telemetry data, post-processing IQ measurements, and exposure information that is collected in tandem with each CFHT observation. Based on this data we build a predictive model of IQ. Through the implementation of a feed-forward mixture density network (MDN, Bishop 1995), we demonstrate that ancillary environmental and operating parameter data are sufficient to predict IQ accurately. Further, we illustrate that, keeping all other settings constant, there exists an optimal configuration of the dome vents that can substantially improve IQ. Our successes here lay the foundation for the development of automated control and scheduling software.

The IQ prediction system we detail in this paper is developed for MegaPrime [2], a wide-field optical system with its mosaic camera MegaCam (Boulade et al. 2003). MegaPrime is one of CFHT's most scientifically productive instruments. Built by CEA in Saclay, France, and deployed in 2003, MegaCam is a wide-field imaging camera. It is used extensively for large surveys covering thousands of square degrees in the sky and ranging in depth from 24 to 28.5 magnitude. MegaCam is placed at the prime focus of CFHT. It includes an image stabilization unit and an auto-focus unit with two

[1] We comment that adaptive optics (Roddier 1988; Beckers 1993) do not solve the problem we are interested in. Adaptive optics are effective only for wavefront errors originating in the upper atmosphere. They cannot correct for ground-layer or dome-induced turbulence. They also most do not work well for wide field imaging systems of the type we consider in this paper.

[2] https://www.cfht.hawaii.edu/Instruments/Imaging/MegaPrime/

Table 1. Brief description of the temperature sensors marked in Figure 1.

Probe label	Description
AIR-4	Air temperature – Caisson, west
SURF-5	Steel temperature – Caisson, east
AIR-6	Air temperature – Upper end, west
AIR-23	Air temperature – Under end, east
AIR-33	Air temperature – Under primary, west
AIR-54	Air temperature – Mirror cell, west
AIR-63	Air temperature – Under primary, south
AIR-65	Air temperature – Inside spigot, north
AIR-67	Air temperature – Under primary, north
GLASS-70	Glass temperature – Under primary, south
AIR-86	Air temperature – Weather tower
AIR-101	Air temperature – MegaPrime exterior

independent guide charge-coupled device (CCD) detectors. MegaCam consists of 40 CCDs, each 2048×4612 pixels in size, for a total of 378 megapixels. The image plane covers a $1^\circ\times1^\circ$ square field of view at a resolution of $0.187''$ (arc-seconds) per pixel. The CFHT archive at the Canadian Astronomy Data Centre (CADC) contains close to 300k Megacam science exposures with 24 filter pass bands. These images have a median IQ of $\sim 0.7''$ in the $r-$band. One of our main results is that, based purely on environmental and observatory operating conditions, we can predict the *effective* MegaPrime IQ (MPIQ for the rest of the paper) to a mean accuracy of about $0.07''$.

We train our models to predict MegaPrime IQ (MPIQ) using CFHT observations dating back to July 23, 2014. While the CFHT data catalogue dates back to 1979, we use data only for the period in which the dome vents have been present. The collected measurements include temperature, wind speed, barometric pressure, telescope altitude and azimuth, and configurations of the dome vents and windscreen. [3] In Table 2 we summarize the environmental sensors, observatory parameters, and miscellaneous features used in this work.

In Gilda et al. (2020), we presented our preliminary results. Here we extend that work significantly and make the following contributions:

(i) We compile and collate several sets of measurements from various environmental sensors, metadata about observatory operating conditions, and measured IQ from MegaCam on CFHT. We curate and combine these sources of data into a single dataset. We publish the curated dataset.

(ii) We use supervised learning algorithms to predict IQ at $0.07''$ accuracy. We present results for a gradient boosting tree algorithm and for a mixture density network (MDN). For the latter we provide a detailed analysis of *feature attributions*, assigning the relative contribution of each input variable to predicting MPIQ.

(iii) The IQ predictions we produce are robust. We perform an uncertainty quantification analysis. Guided by a robust variational autoencoder (RVAE) that models the density of the data set, we identify non-representative configurations of our sensors.

(iv) We use our MDN to find the optimal vent configurations that would have resulted in the lowest IQ. We use these predictions to estimate the annual increase in science return and scientific observations. We find the improvement to be $\sim 15\%$. This improvement results from increased observational efficiency at CFHT, in particular minimizing the observation times for hypothetical r-band targets

[3] We have made our data set publicly available at https://www.cfht.hawaii.edu/en/science/ImageQuality2020/.

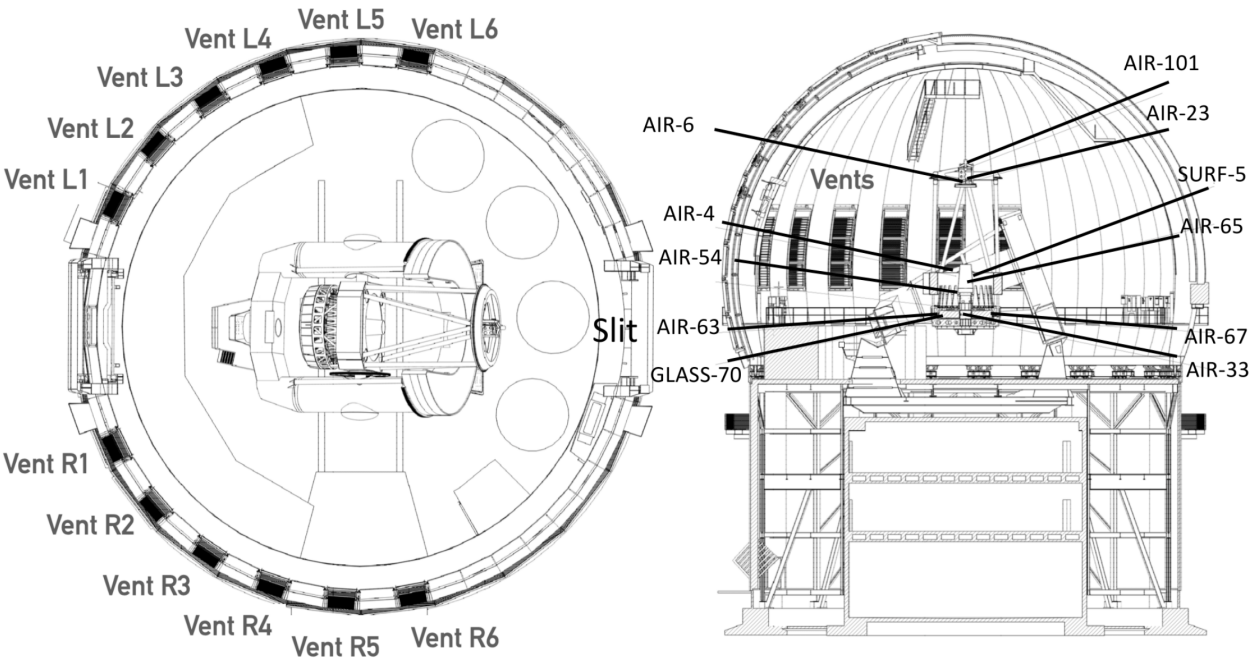


Figure 1. A schematic of the CFHT; top-view and profile. The twelve actionable dome vents are marked. Important thermal sensors identified in past works (see Section 2) are highlighted. These sensors are detailed in Table 1.

Table 2. Data fields in the MegaCam dataset.

Parameter	Units	#Features	Range	Description
Environmental				
Temperature	°C	57	[-8,20] / [-200,850]	Temperature values from sensors in and around the dome. Three sensors are placed within the dome. The rest are external.
Wind speed	knots	1	[0,35]	Wind speed at the weather tower.
Wind azimuth	NONE	2	[-1,1]	Sine and cosine of wind azimuth with respect to true north.
Humidity	%	2	[1.4,100]	Measured both at the top of the observatory dome, and at the weather tower.
Dew point	%	2	[1.4,100]	Measured both in the basement of the observatory building, and at the telescope mirror cell (near GLASS 70 in Figure 1)
Barometric pressure	mm of Hg	1	[607,626]	Atmospheric pressure measured on the fourth floor of the observatory building.
MPIQ	” of Hg	1	[0.35,2.36]	Measured seeing from MegaCam/MegaPrime.
Observatory				
Vents	NONE; NONE; NONE	36	{0,1}; [0,1]; [0,1]	For each sample, we have three types of vent values: vent configuration (‘OPEN’ or ‘CLOSE’), and Sine and cosine of vent_{AZ}
Dome azimuth	NONE	2		Sine and cosine of the angle of the slit-center from true North.
Pointing altitude	NONE	1	[0.15,1]	Sine of the angle of the telescope focus from the horizontal.
Pointing azimuth	NONE	2	[-1,1]	Sine and cosine of angle of the telescope focus from true north.
Wind screen position	NONE	2	[0,1]	Fraction that the wind screen is open. (The wind screen is located at the ‘Slit’ position in the left of Figure 1.)
Central Wavelength	nm	1	[354,1170]	Central wavelengths of each of the 22 filters.
Dome Az – Pointing Az	NONE	2	[-1,1]	Sine and cosine of difference between dome and pointing azimuths.
Dome Az – Wind Az	NONE	2	[-1,1]	Sine and cosine of difference between dome and wind azimuths.
Pointing Az – Wind Az	NONE	2	[-1,1]	Sine and cosine of difference between wind and pointing azimuths.
Other				
Exposure time	seconds	1	[30,1800]	Observation time per sample.
Observation Time	NONE	4	[-1,1]	Sine and cosine of $\frac{\text{hour_of_day}}{23}$ and $\frac{\text{week_of_year}}{51}$.

of the 25th magnitude to achieve an SNR of 10; these figures are representative of deep observations of faint targets of large imaging programs at CFHT like the Canada France Imaging Survey [4].

The rest of the paper is structured as follows. In Section 2 we discuss relevant previous work. In Section 3 we explore in depth the various sources of input data and the processing pipeline we implement to collate and convert the data sources into the final usable dataset. In Section 4 we describe in detail our methodology, including attributes of our gradient boosting tree and neural network methods, feature importance method, and our predictions for best vent configurations. In Section 5 we present our results. We conclude in Section 6. To help keep our focus on astronomy, some supporting figures that help detail our machine-learning implementations are deferred to Appendix A.

2 RELATED WORK

The summit of Maunakea was selected as the site for CFHT due to its excellent astronomical observing properties: low turbulence, low humidity, low infrared absorption and scattering, excellent weather, clear nights. Image quality, or "seeing", is quantified using the full-width half-max (FWHM) measure. FWHM, expressed in arc-seconds ($''$), is calculated as the ratio of the width of the support of a distribution, measured at half the peak amplitude of the distribution, to half the peak-amplitude value; smaller FWHM is better. For example, the FWHM of the Gaussian distribution is $2\sqrt{2\log 2}\sigma$, roughly 2.4 times the standard deviation σ. In our application, FWHM operationally quantifies the degree of blurring of uncrowded and unsaturated images of point sources (such as a star or a quasar) on the central CCDs of a MegaCam frame. The FWHM measured this way, referred to as image quality or IQ, is an aggregate of multiple sources.[5] The main contributors to FWHM / IQ are: imperfections in the optics ($\mathrm{IQ}_{\mathrm{Optics}}$), turbulence induced by the dome ($\mathrm{IQ}_{\mathrm{Dome}}$), and atmospheric turbulence ($\mathrm{IQ}_{\mathrm{Atmospheric}}$). These contributions are well-modeled as being independent and as combining to form the measured IQ ($\mathrm{IQ}_{\mathrm{Measured}}$) according to

$$\mathrm{IQ}^{5/3}_{\mathrm{Measured}} = \mathrm{IQ}^{5/3}_{\mathrm{Optics}} + \mathrm{IQ}^{5/3}_{\mathrm{Dome}} + \mathrm{IQ}^{5/3}_{\mathrm{Atmospheric}}. \quad (1)$$

If the contributions were modeled by a Gaussian distribution, the exponents in Equation (1) would be $2 = 6/3$ (because variances of independent Gausian random variables add). The $5/3^{\mathrm{rd}}$ power is due to the spectrum of turbulence which was characterized by Kolmogorov in 1941 (V. I. Tartarskiĭ & R. A. Silverman (ed.) 1961). We note that while the contribution of the optics is not due to turbulence, we still use the power of $5/3$ in our model for consistency. Finally, we note that of the three contributors we can only influence $\mathrm{IQ}_{\mathrm{Dome}}$ through actuation of various observatory controls.

While the mean free atmosphere ($\mathrm{IQ}_{\mathrm{Atmosphere}}$) seeing on Maunakea is estimated to be about $0.43''$ (Salmon et al. 2009), in practice, the IQ realized at CFHT is usually worse (i.e. the seeing is higher). Through 40 years of effort by the CFHT staff and consortium scientists, $\mathrm{IQ}_{\mathrm{Measured}}$ has steadily decreased, from early values of $2''$ or greater to its current median value of around $0.71''$. Removal of $\mathrm{IQ}_{\mathrm{Optics}}$ further reduces this figure to $0.55''$ (see Figure 2, Equation 2, and Section 3.3).[6] In the remainder of this section we discuss prior efforts to quantify IQ and to reduce IQ. Later, in Section 3 when we discuss our data sources, we return to (1) and step through a number of sources of variation in observing conditions (e.g., wavelength of observation, elevation of observation) that we correct to produce a normalized dataset in which IQ measurements from distinct observations can be directly compared.

Early published efforts to quantify and reduce the IQ at CFHT (e.g. Racine (1984)) detail campaigns to minimize turbulence inside and around the dome, including analysis and measurements of the opto-mechanical imperfections of the telescope. The team led by René Racine estimated that if the in-dome turbulence was corrected and the telescope imperfections were removed, the natural Maunakea seeing would offer images with IQ below $0.4''$ FWHM on one-quarter of the nights. Later efforts (Racine et al. 1991) used data from the (then) new HRCam, a high-resolution imaging camera at the prime focus of CFHT, to develop a large and homogeneous IQ data set. They correlated their IQ data with thermal sensor data, through which they were able to identify and quantify "local seeing" effects. Their main findings, relevant to our work, are listed below.

(i) The contribution of mirror optics $\mathrm{IQ}_{\mathrm{Optics}}$ amounts to about $0.4''(\Delta T)^{6/5}$ where ΔT is the temperature difference between the primary mirror and the surrounding air.
(ii) The dome contribution $\mathrm{IQ}_{\mathrm{Dome}}$ amounts to about $0.1''(\Delta T)^{6/5}$ where ΔT is the temperature difference between the air inside and outside of the dome.
(iii) The median natural atmospheric seeing at the CFHT site $\mathrm{IQ}_{\mathrm{Atmospheric}}$ is $0.43'' \pm 0.05''$. The 10th and 90th percentiles are roughly $0.25''$ and $0.7''$.

More recent follow-up work is presented in Salmon et al. (2009). The authors correlate measured IQ using the (then) new MegaCam with temperature measurements. They analyze 36,520 MegaCam exposures made in the u, g, r, i, and z-bands in the three year period between August 2005 and August 2008. They find strong dependencies of the measured IQ on temperature gradients. Furthermore, in Table 4 of Salmon et al. (2009) the authors categorize important factors that contribute to the seeing – atmosphere, dome, optics – and provide estimates of their respective contribution. As the authors discuss, these estimates update the findings of Racine et al. (1991). The most significant findings of Salmon et al. (2009) can be summarized as follows.

(i) The orientation of the dome slit with respect to the wind direction has important effects on IQ.
(ii) The median dome induced seeing $\mathrm{IQ}_{\mathrm{Dome}}$ before the installation of the vents in 2013 was $0.43''$.
(iii) The seeing contribution from optics and opto-mechanical imperfections $\mathrm{IQ}_{\mathrm{Optics}}$ varied from $0.46''$ in the u-band to $0.28''$ in the i-band.
(iv) Atmospheric seeing $\mathrm{IQ}_{\mathrm{Atmospheric}}$ at the CFHT site at a wavelength of 500 nm and an elevation of 17m above ground was measured using a separate imager. The median $\mathrm{IQ}_{\mathrm{Atmospheric}}$ measured was $0.55''$. This estimate of atmospheric seeing is independent of effects related to the dome and the optics.

The culminating result of these studies that analyzed the delivered IQ was the December 2012 installation, and July 2014 initial use, of the 12 dome vents depicted Figure 1. Since their installation,

[4] https://www.cfht.hawaii.edu/Science/CFIS/
[5] Note that larger FWHM $\implies$ higher "seeing" $\implies$ poorer image quality (IQ) $\implies$ more arc-sec. So, lower FWHM which equates to a better IQ (fewer arc-sec) is desired.
[6] https://www.cfht.hawaii.edu/Science/CFHLS/T0007/T0007-docsu11.html

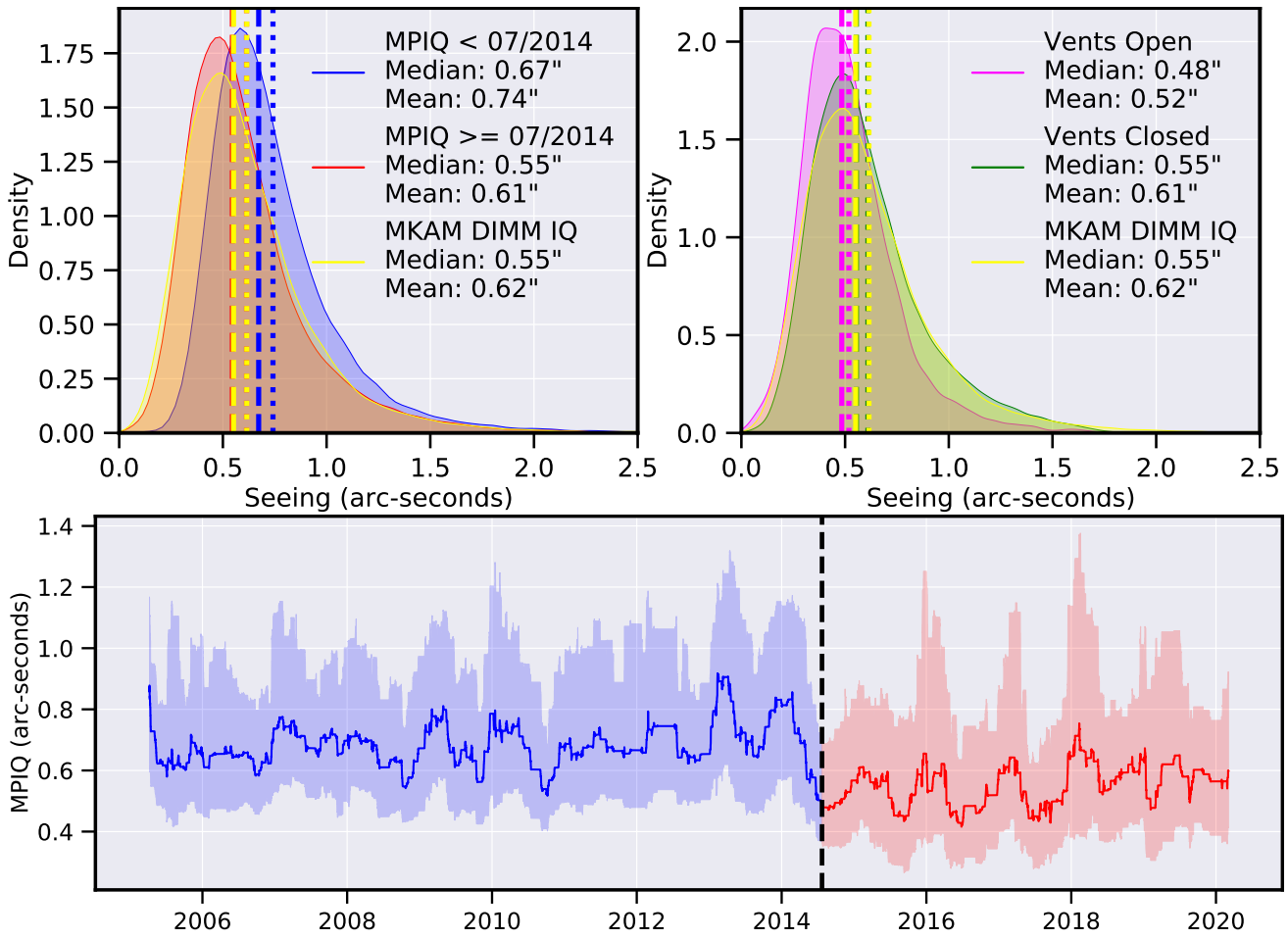


Figure 2. Seeing evolution and distribution. **Upper left:** Distribution of seeing measured by the MKAM Differential Image Motion Monitor (DIMM) and effective IQ measured at MegaPrime since 2010. Effective MPIQ is the measured IQ less the contributions from optics – see Equation (1) and Section 3.2. Both the MKAM and the MPIQ curves peak at $\sim 0.55''$; the former contains relatively higher seeing contribution from the ground layer, while the latter includes contribution from the dome itself (see Section 2 for detailed discussion). Both peaks are higher than the best possible seeing at the site of $\sim 0.43''$. **Upper right:** Distribution of seeing *after* the installation of the vents as a function of vent configuation: *all-open* or *all-closed*. Observe the statistics of the former are much better than those of the latter. We also plot the MKAM histogram, which is basically unchanged from prior to the installation of the vents. **Lower:** Quarterly averaged MegaPrime IQ of the CFHT. The wiggly curve of decaying mean and oscillation amplitude is a model that peaks in midwinter, when the outside air temperature tends to be colder than the dome air. The drop in July of 2014 corresponds to when the vents first started to be used; this is why in the data input to our models we do not use samples from before this month.

CFHT operators have kept all 12 vents completely open as often as possible, barring conditions of mechanical failure and strong winds. As already mentioned, this allows faster venting of internal air and equalization of internal and external temperatures.[7] The vent-related improvement in seeing has been dramatic, with median $\mathrm{IQ}_{\mathrm{Measured}}$ improving from about $0.67''$ to $0.55''$ [8].

In order to have an external, regularly sampled seeing reference, we use the Maunakea Atmospheric Monitor (MKAM, Skidmore et al. 2009; Tokovinin et al. 2005). This telescope, dedicated to seeing monitoring, is mounted on top of a weather tower just outside of CFHT. It has a composite instrument, including a Multi Aperture Scintillation Sensor (MASS) and a Differential Image Motion Monitor (DIMM). We only use data from the latter as the former is insensitive to the lower layers of turbulence. DIMMs measure seeing by computing the variance of the relative motion of the images formed by two separate sub-apertures, therefore probing the curvature of the wave-front. This variance can be directly related to the the full-width at half-max (FWHM) of the point spread function (PSF) in long exposures given the wavelength of observation and the sub-apertures diameter (see e.g. Sarazin & Roddier 1990). While MKAM measurements are free of the CFHT dome contribution to seeing, they are also sensitive to part of the ground layer contributions to seeing that are not seen by the CFHT instruments, partly due to the lower altitude of the weather tower as compared to the CFHT aperture, and to localized differences in the summit turbulence in the first few meters above ground. MKAM thus serves as a slightly noisy seeing reference for Megacam, free of CFHT dome seeing contributions.

In the left sub-plot of Figure 2, we plot histograms of the *corrected* seeing values from MegaCam both before (starting February 2nd 2002) and after July 7th 2014, when the vents started being used; the latter is the the start date of our data set used in the

[7] http://www.cfht.hawaii.edu/AnnualReports/AR2012.pdf
[8] http://www.cfht.hawaii.edu/AnnualReports/AR2014.pdf

remainder of this work). We compare these with the seeing distribution from MKAM and MegaCam for observations since 2002. Corrected seeing removes the contribution of the telescope optics from the measured seeing, and is defined in Equation 2. In the top-right plot in Figure 2, "Vents Open" refers to samples where the 12 vents are either all open, or at most one of them is closed, whereas "Vents Closed" incorporates samples where all 12 vents are closed. As can be seen, the introduction of the vents has reduced the median MPIQ by $0.21'' = \left((0.55^{5/3} - 0.48^{5/3})^{3/5}\right)$, and the mean MPIQ by $0.27'' = \left((0.61^{5/3} - 0.51^{5/3})^{3/5}\right)$. However, there is still money on the table – the estimated free-air, observatory-free IQ at the CFHT site is estimated to be $\sim 0.43 - 0.44''$ (Salmon et al. 2009). This means that there is still a possible improvment in median IQ of $0.15'' = \left((0.48^{5/3} - 0.435^{5/3})^{3/5}\right)$, and a mean IQ of $0.23'' = \left((0.52^{5/3} - 0.435^{5/3})^{3/5}\right)$. This range of improvement was independently verified by another CFHT team in 2018 (Racine et al. 2018). They found that, when open, the dome vents on average reduce IQ by $0.37''$. While this number is significantly larger their either of our estimates of $0.21''$ or $0.27''$, their estimate of degradation of IQ of about $0.20''$, caused by residual eddies induced by thermal differences in the dome, closely matches our own. It is precisely this residual part of IQ_{Dome} that we aim to capture in our work. We later show in Figures 10c and 10d, and in Section 5.3 that our models are indeed able to capture these improvements.

We remind the reader, as mentioned above, that CFHT operators have kept all 12 vents completely open as much as possible. They have chosen this manner of operation as they had no basis upon which to choose a more varied configuration of the dome vents. Although, we note that fluid flow modeling conducted during the vent design process predicted that intermediate settings (i.e., neither all-optimal nor all-closed) would optimally reduce internal turbulence (Baril et al. 2012). By tuning the dome vent configurations, based on current environmental conditions, to a setting between all-open and all-closed, we aim to reduce this residual.

From a programmatic perspective, our work is a natural extension of Racine (1984); Racine et al. (1991); Salmon et al. (2009); Racine et al. (2018). While these prior investigations correlated IQ with measurements of temperature gradients, our work tries to relate all of the metrics (not solely the temperature metrics) with the $\text{IQ}_{\text{Measured}}$ through the application of advanced machine learning techniques. Further, rather than only establishing correlation, we also seek to understand whether by actuating the dome parameters under our control we can improve the delivered $\text{IQ}_{\text{Measured}}$. Recent work at the Paranal Observatory by Milli et al. (2019) similarly collected 4 years of sensor data, and trained random forest and neural networks to model and forecast over the short term (< 2 hrs) the DIMM seeing and the MASS-DIMM atmospheric coherence time and ground layer fraction. Their early results demonstrate good promise, especially for scheduling adaptive optics instruments.

Finally, we mention recent work (Lyman et al. 2020) by the Maunakea Weather Center (MKWC) which takes a macro approach to predict $\text{IQ}_{\text{Atmospheric}}$. The authors tap into large meteorological modeling models. They start from the NCEP/Global Forecasting System (GFS) which outputs a 3D-grid analyses for standard operational meteorological fields: pressure, wind, temperature, precipitable water, and relative humidity. Coupling these predictions with advanced analytics and decades of MKAM DIMM seeing data, Lyman et al. (2020) predict the free air contribution ($\text{IQ}_{\text{Atmospheric}}$) to seeing on the mountain. Their work is complementary to ours in that we take in our local sensor measurements to predict (and reduce) the effect of IQ_{Dome} on $\text{IQ}_{\text{Measured}}$, while Lyman et al. (2020) directly predict $\text{IQ}_{\text{Atmospheric}}$. In the long term these two models can be combined to yield improved seeing estimates, forecasts, and decisions.

3 DATA

In this section we discuss how we curated and prepared the data for use in our models. As mentioned already, our efforts began with almost a decade's worth of sensor measurements archived at CFHT, together with IQ measured on the MegaCam exposures retrieved from the Canadian Astronomy Data Center (CADC) web services. At the start pertinent variables were spread across multiple data sets, sensor measurements were missing due to sensor failures, data records contained errant values due to mis-calibrated data reduction pipelines. We therefore spent substantial effort cleaning the data. In Section 3.1 we discuss the various data sources that we collate to form our final data set. We then discuss our data cleaning and feature engineering procedures in Section 3.2 and Section 3.3.

3.1 Data sources

Our first step in data collection was to build a data archive that contains one record per MegaCam exposure. In the remainder of this paper, we refer to each exposure and its associated sensor measurements interchangeably as a "sample", a "record", or an "observation". Each record contains three distinct types of predictive variables:

(i) *Observatory parameters*. These can be divided into operating (controllable) and non-operating (fixed) parameters. The former include the configurations of the twelve dome vents (open or closed), and the windscreen setting (degrees open). These are examples of the variables that we can adjust the settings of in real time. The non-operating features include measurements of the telescope altitude and azimuth (which correspond to pointing of the astronomical object being observed) and the central wavelength of the observing filter.

(ii) *Environmental parameters*. These include exposure-averaged wind speed, wind direction, barometric pressure, and temperature values at various points both inside and outside the observatory.

(iii) *Ancillary parameters*. Each exposure come with metadata. Relevant to our work are the date and time of the observation and the length of exposure. All predictive variables have been summarized in Table 2. The median time of each exposure is ~150 seconds, while the median time between two consecutive exposures made on the same night is ~240 seconds[9].

In total, there are 160,341 observations, and 86 variables (*including* image quality) are provided with each exposure. The records span February 2005 to March 2020. An overview of the data is provided in Table 3, where we note the expanded number of features created using *feature engineering*, which we expound upon below.

[9] We emphasize that in this work we forego temporal dependencies and treat all exposures as independent. We provide the time between exposures for the sake of context.

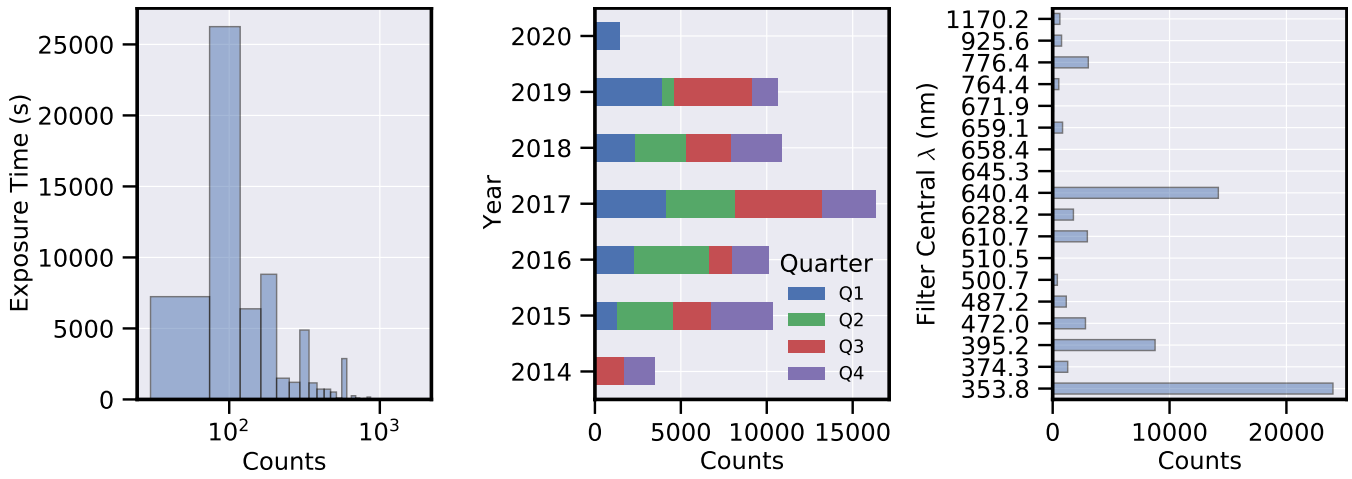


Figure 3. Left: Histogram of exposure times of the ~60,000 samples/exposures used in this paper. **Middle:** Collected observations per year, broken down by quarter. **Right:** Same observations group by filter as represented by the filter central wavelength.

3.2 Data Cleaning

We now list the data cleaning we performed. In short these included the removal of data records corresponding to (i) non-sidereal targets, (ii) data records associated with too-short or too-long exposures, (iii) data records associated with IQ estimates deemed unrealistic, and (iv) data records containing missing or errant data values.

(i) *Non-sidereal:* We remove moving, non-sidereal, targets. The IQ measurements for these data records are not valid as the data pipeline that calculates IQ assumes sidereal observation. Therefore these records are not appropriate for training. We note that as part of the configuration data recorded along with each observation the astronomer specifies whether the observation is sidereal or not. Hence, these data records are easy to remove.

(ii) *Non-trustworthy IQ estimates:* We remove MegaCam exposures associated with IQ estimates of less than $0.15''$ or greater than $2''$. Such IQ numbers are deemed unrealistic. It is believed that an IQ of $\sim 0.2''$ is the best possible at Maunakea. Anything below this is deemed to result from an erroneous calculation when converting from the raw exposure data. On the other hand, IQ $> 2''$ is too large for useful science.

(iii) *Missing and errant measurements:* Not all sensor measurements are available at all times of the exposure. We refer to these as "missing data". As is tabulated in the first row of Table 3, prior to considering missing data, our cleaned data set (cleaned of non-sidereal and non-trustworthy) contains 120 features (86 original + rest engineered features + 1 MPIQ, see Section 3.3) and 160,341 samples. Of these, just under 100,000 samples *do not* contain all measurements; we specify the fraction of missing measurements in the last column of Table 3. We refer to the dataset with *all* samples as $\mathcal{D}_{\mathcal{F}_\mathcal{S},\mathcal{S}_\mathcal{L}}$, i.e., $\mathcal{D}$ata set with a $\mathcal{S}$mall number of $\mathcal{F}$eatures, and a $\mathcal{L}$arge number of $\mathcal{S}$amples. By removing those samples that contain at least one missing feature, we obtain $\mathcal{D}_{\mathcal{F}_\mathcal{S},\mathcal{S}_\mathcal{S}}$: $\mathcal{D}$ata set with a $\mathcal{S}$mall number of $\mathcal{F}$eatures, and a $\mathcal{S}$mall number of $\mathcal{S}$ samples. This latter dataset consists of 63,082 samples (second and third rows in Table 3. In this paper, we train our models on $\mathcal{D}_{\mathcal{F}_\mathcal{S},\mathcal{S}_\mathcal{S}}$, since feed-forward neural networks cannot handle missing values without non-trivial modifications. In future work, we will use a variational autoencoder capable of imputing missing values (Collier et al. 2020) to enable us to leverage the larger dataset, $\mathcal{D}_{\mathcal{F}_\mathcal{L},\mathcal{S}_\mathcal{L}}$: $\mathcal{D}$ata set with a $\mathcal{L}$arge number of $\mathcal{F}$eatures, and a $\mathcal{L}$arge number of $\mathcal{S}$amples.

Table 3. Summary statistics of data sets described in Section 3.2. '#Original Features' includes the MegaPrime Image Quality (MPIQ), while '#Engineered Features' are additional hand-crafted ones added to enhance the predictive capability of our models (see Section 3.3 for details). However for the remainder of this paper, we use 'features' to refer to the union of original and engineered features less the MPIQ column: these are predictive, independent variables. Similarly, going forward MPIQ – the dependent variable – is referred to as the 'target'.

Dataset Identifier	#Samples	#Original Features	#Engineering Features	Percentage Missing
$\mathcal{D}_{\mathcal{F}_\mathcal{S},\mathcal{S}_\mathcal{L}}$	160,341	86	34	62%
$\mathcal{D}_{\mathcal{F}_\mathcal{S},\mathcal{S}_\mathcal{S}}$	63,082	86	34	0%
$\mathcal{D}_{\mathcal{F}_\mathcal{L},\mathcal{S}_\mathcal{S}}$	63,082	86	1115	0%
$\mathcal{D}_{\mathcal{F}_\mathcal{L},\mathcal{S}_\mathcal{L}}$	160,341	86	1115	62%

3.3 Feature Engineering

Feature engineering is the process of modifying existing features, using either domain expertise, statistical analysis, or intuition derived from scientific expertise. The goal is to create predictive variables that are more easily understood by an ML algorithm. We now describe the feature engineering we performed.

(i) *Optics IQ correction:* We remove the fixed, but wavelength dependent, contributions of the telescope optics to IQ, $\mathrm{IQ}_{\mathrm{Optics}}$. These corrections are based on work by Salmon et al. (2009), and range from $0.31''$ in the i-band to $0.53''$ in the u-band (Racine et al. 2018); cf. second column of Table 4. After removing the contribution of optics, we are left with a convolution of dome seeing and atmospheric seeing. This is because dome seeing, referred here to as $\mathrm{IQ}_{\mathrm{Dome}}$, is enmeshed with $\mathrm{IQ}_{\mathrm{Measured}}$ in a complicated way that does not lend itself to easy separation; the relationship between these is governed by Equation (2), a rearranged version of Equation (1):

$$\mathrm{IQ}^{5/3}_{\mathrm{Atmospheric}} + \mathrm{IQ}^{5/3}_{\mathrm{Dome}} = \mathrm{IQ}^{5/3}_{\mathrm{Measured}} - \mathrm{IQ}^{5/3}_{\mathrm{Optics}},$$
$$\mathrm{IQ}'_{\mathrm{Atmospheric}} = \left(\mathrm{IQ}^{5/3}_{\mathrm{Measured}} - \mathrm{IQ}^{5/3}_{\mathrm{Optics}}\right)^{3/5} \quad (2)$$

At the risk of being redundant with information presented towards the tail-end of Section 2, we remind the readers that Racine et al. (2018) and Salmon et al. (2009) estimate $\mathrm{IQ}_{\mathrm{Atmospheric}}$ to be in

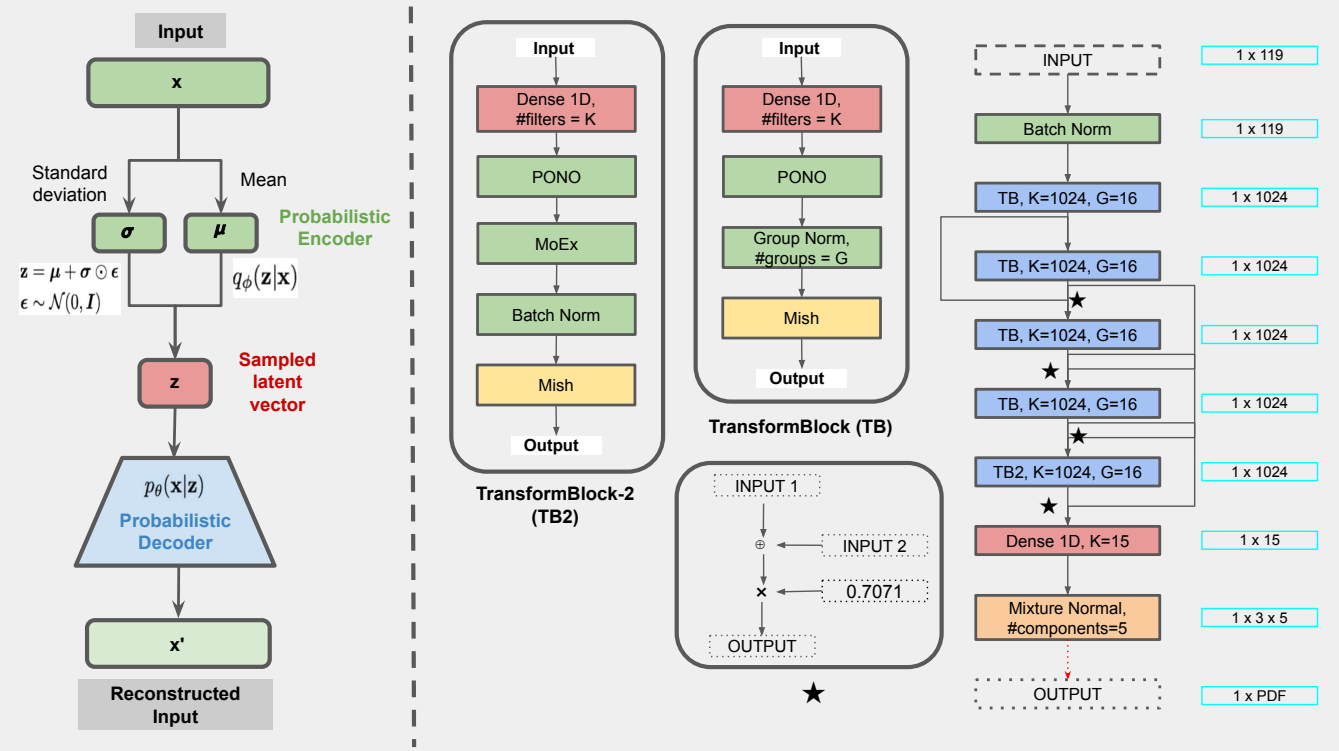


Figure 4. The architectures of the two networks used in this study. To the **left** of the dashed vertical line, we show the overview of a variational autoencoder, but note that the *robust* VAE (i.e. RVAE) used in this work uses a special reconstruction loss (comparing **x** and **x'**) which is not depicted in this cartoon. On the **right** we show a dense feed-forward network with skip connections, mish activations (Misra 2019), positional normalization (Li et al. 2019), batch normalization (Ioffe & Szegedy 2015), group normalization (Wu & He 2018), and momentum exchange (Li et al. 2020) augmentation layers. This mixture density network (MDN) has 5 components. Near the right edge of the figure, we indicate in cyan colored rectangles the output shapes after an input sample has passed through each layer.

Table 4. IQ_{Optics} for different bands, calculated according to the prescription of Salmon et al. (2009). The average seeing across all bands is about 0.33″, as noted in Table 4 of Salmon et al. (2009).

Filter	Central λ (nm)	IQ_{Optics} (″)
Ha.MP7605	645.3	0.284
HaOFF.MP7604	658.4	0.280
Ha.MP9603	659.1	0.280
HaOFF.MP9604	671.9	0.276
TiO.MP7701	777.7	0.260
CN.MP7803	812	0.260
u.MP9301	374.3	0.441
u.MP9302	353.8	0.459
CaHK.MP9303	395.2	0.424
g.MP9401	487.2	0.358
g.MP9402	472	0.368
OIII.MP7504	487.2	0.358
OIII.MP9501	487.2	0.358
OIIIOFF.MP9502	500.7	0.350
r.MP9601	628.2	0.290
r.MP9602	640.4	0.285
gri.MP9605	610.68	0.296
i.MP9701	777.6	0.261
i.MP9702	764.4	0.261
i.MP9703	776.4	0.261
z.MP9801	1170.2	0.397
z.MP9901	925.6	0.276

the range of 0.43″ to 0.45″, and $IQ'_{Atmospheric}$ to be about 0.55″. They demonstrate that opening all 12 vents completely allows one to reduce $IQ'_{Atmospheric}$ to about 0.51″, which leaves a residual median 0.20″ $\left((0.51^{5/3} - 0.43^{5/3})^{3/5}\right)$ on the table, which is what we aim to capture in this paper. These numbers also agree with our own calculations, as described in Section 2 and visualized in the top two sub-figures of Figure 2. Our argument, introduced in Section 1 and expounded upon in Section 1, is that for any given observation, there is an optimal set of vent configuration, somewhere between all-open and all-closed, that allows us to bite into this 0.20″ residual IQ_{Dome}.

(ii) *Wavelength IQ correction:* Each MegaCam exposure is taken using one of 22 band-pass filters. The right-hand subfigure in Figure 3 plots a histogram of observations across bands. The use of the filters results in a wavelength-dependent IQ variation. To make IQ measurements consistent we scale IQ to a common wavelength of 500nm. The formula for the scaling is provided in Equation (3), which we present in conjunction with a zenith angle correction, discussed next.

(iii) *Zenith angle correction:* IQ is also affected by the amount of atmosphere through which the observation is made. The contribution of air mass is, to first degree, predictable, and can be removed together with the wavelength correction via (3), where z is the zenith angle in degrees and λ is the central wavelength of a given filter in nm.

$$\mathrm{IQ}_{\mathrm{Corrected}} = \mathrm{IQ}'_{\mathrm{Atmospheric}} \times \left(\frac{\lambda}{500}\right)^{1/5} \times (\cos z)^{3/5} \quad (3)$$

(iv) *Cyclic encoding of time-of-day and day-of-year:* Every observation has an associated time stamp, indicating the beginning of image acquisition. Using this 'timestamp' feature, we derive two time-features, the hour-of-day and the day-of-year. These features better capture latent cyclical relationship between weather events and IQ. We represent each of these two features into a pair cyclical 'sinusoidal' and 'co-sinusoidal' component. For example, for the day-of-week feature values – which can range from 0 to 6 – we encode it as day-of-week-sine and day-of-week-cosine. These can each respectively take on values from $\sin(0 \times 180°/6)$, to $\sin(6 \times 180°/6)$, and $\cos(0 \times 180°/6)$, to $\cos(6 \times 180°/6)$. In this way we replace the timestamp feature with four new, and more easily digestible features.

(v) *Cyclic encoding of azimuth:* Similar to the temporal information, we cyclically encode the telescope azimuth, splitting it into two features. We note that since the altitude of observation ranges from 0 to 90 degrees, and is not cyclical in nature, we leave that feature unmodified.

(vi) *Temperature differences features:* As argued in our discussion, and evidenced by the prior work, temperature differences are the prime source of turbulence. In recognition of this key generative process, we engineer new temperature features that consist of the pairwise differences of existing temperature measurements.

We note that, given sufficient data, a deep enough neural network should be able to discover that temperature differences are important features. We engineer in such features as, from our knowledge of physics, we understand temperature difference are important and providing them explicitly to the network eases the inference task faced by the network. In addition, unlike a neural network, the boosted-tree model that we use for comparative analysis is, by design, unable to create new features. The boosted-tree therefore benefits quite significantly from increased feature representation.

We implement two different flavors of engineering here. First, for every temperature feature in our two data sets of 160,341 and 63,082 samples, we subtract it from every other temperature feature. We calculate the Spearman correlation of these newly generated features with the MPIQ values. We then rank them by magnitude in descending order and pick the top three features. This increases our original 85 input features to 119, and this is how we get $\mathcal{D}_{\mathcal{F}_S,\mathcal{S}_\mathcal{L}}$ and $\mathcal{D}_{\mathcal{F}_S,\mathcal{S}_S}$. For the second variation, we do not pick the top 3, but retain all the newly generated temperature-difference features. This increases the number of features from 86 to 1119. This is how we get $\mathcal{D}_{\mathcal{F}_\mathcal{L},\mathcal{S}_\mathcal{L}}$ and $\mathcal{D}_{\mathcal{F}_\mathcal{L},\mathcal{S}_S}$. This is summarized in Table 3. As a reminder, in this work, we only use $\mathcal{D}_{\mathcal{F}_S,\mathcal{S}_S}$; empirical results showed that our neural networks' performance did not significantly improve by using $\mathcal{D}_{\mathcal{F}_\mathcal{L},\mathcal{S}_S}$.

4 METHODOLOGY

The raw sensor data is a collection of time series and ultimately it would best to model the multiple sensors in their native data structure. In the analysis we perform in this paper, we compiled the sensor data into a large table to ease exploration, consisting of heterogeneous and categorical data. The heterogeneity is caused by the wide variety of sensors (wind speed, temperature, telescope pointing) each recorded in specific units. Categorical features emerged because certain measurements values were binned. For instance, due to the unreliability of wind speed measurement, we have binned these values – wind speed below 5 knots, 5-10 knots, etc. Similarly, for simplicity, each of the twelve vents have been encoded into either completely open or completely closed. These characteristics induce a discontinuous feature space. Our training data set is thus tabular in nature. At hand with our curated data set, we are equipped to work on our two objectives: making accurate predictions of MegaPrime Image Quality, and use our predictor to, on a per-sample basis, explore the importance of each feature on IQ.

Decision tree-based models (Quinlan 1986) and their popular derivatives such as random forests (Breiman 2001), and gradient boosted trees (Friedman 2001) are well-matched to tabular data. and often are the best performers. Tree-based models select and combine features greedily to whittle down the list of pertinent features to include only the most predictive ones. Feature sparsity and missing data is naturally accommodated by tree models, they simply do not include feature cells containing such values in their splits. We show below our implementation of a variant of gradient boosted tree with uncertainty quantification, and feature exploration.

However tree-based models require the human process of feature engineering and are known (e.g. Bengio et al. (2010)) to poorly generalize. In contrast to tree-based models, deep neural networks (DNNs) are powerful feature pre-processors. Using back-propagation, they learn a fine-tuned hierarchical representation of data by mapping input features to the output label(s). This allows us to shift our focus from feature engineering to fine-tuning the architecture, designing better loss functions, and generally experimenting with the mechanics of our neural network. In reported comparison cases, DNNs yield improved performance with larger sized datasets (Haldar et al. 2019). As we will show, our neural network implementation, with the feature engineering steps described above, performs better than the alternative tree-based boosted model. We therefore deepen our analysis of the deep neural networks further: we quantify its probabilistic predictions, and we attempt to model the feature space.

4.1 Probabilistic Predictions with a Mixture Density Network

Mixture density networks (MDNs) are composed of a neural network, the output of which are the parameters of a mixture model (Bishop 1995). They are of interest here because the relationship between the feature vectors $\mathbf{x}$ and target labels $\mathbf{y}$ can be thought of stochastic nature. Therefore, MDNs express the probability distribution parameters of MPIQ as a function of the input sensor features. In a one-dimensional mixture model, the overall probability density function (PDF) is a weighted sum of M individual PDF $p^m_\theta(\mathbf{y}|\mathbf{x})$ parameterized by a neural network of parameters θ [10]:

$$p_\theta(\mathbf{y}|\mathbf{x}) = \sum_{m=1}^{M} \alpha^m_\theta(\mathbf{x}) p^m_\theta(\mathbf{y}|\mathbf{x}) \quad \text{with} \quad \sum_{m=1}^{M} \alpha^m_\theta(\mathbf{x}) = 1.$$

Under the assumption of N independent samples $\mathbf{x}$ from the features distribution, and the corresponding conditional samples $\mathbf{y}$ of MPIQ, we minimize over the negative log-likelihood of the density mixture to obtain the neural network weights:

$$\theta^* = \underset{\theta}{\operatorname{argmin}} -\frac{1}{N} \sum_{n=1}^{N} \log p_\theta(\mathbf{y}_n|\mathbf{x}_n) \quad (4)$$

[10] In their initial form Bishop (1995), MDNs used a Gaussian mixture model (GMM). They can easily be generalized to other distributions.

To train the neural network we take as the network input the data record of sensor readings, observatory operating conditions, etc. The network outputs are the (per-vector) mixture model parameters modeling the MPIQ conditional distribution, implicitly parameterized by the neural network. In our experiments, we use β distributions, and set $M = 5$ as it gave sensible results.

4.2 Complementary Predictions and Interpretation with Gradient Boosted Decision Trees

We complement the MDN IQ predictions by another algorithm to secure our results: a gradient boosted decision tree (GBDT) to predict IQ from the sensor data. This is in fact one of the main reason so much of feature engineering was performed on the sensor data. A set of consecutive decision trees is fit where each successive model is fit to obtain less overall residuals than the previous ones by weighting more the larger sample residuals. Once converged, we can obtain a final predictions from the trained boosted tree as the weighted mean of all models. The optimization can be performed with gradient descent. Several implementations of this popular algorithm exist and we selected the CATBOOST[11] one for our modelling, with a loss optimized for both the mean and the variance of the predictions. We first perform nested cross-validation as for the MDN, obtain the best hyper-parameters. We then train multiple (10) GBDT models with the same hyper-parameters with a stochastic optimization, each with a different initialization of the model parameters. Aleatoric and epistemic uncertainties are estimated with a simple ensemble averaging method (Malinin et al. 2021) of each model predictions. We show our results in Figure 6, and discuss them in detail in Section 5.

4.3 Density Estimation with a Robust Variational Autoencoder

An autoencoder (Hinton & Salakhutdinov 2006) is a neural network that takes high dimensional input data, encode it into a common efficient representation (usually of lower dimension), and then recreates a full-dimensional approximation at the other end. Through its mapping of the input to a smaller or sparser, and more manageable, but information-dense, latent vector, the autoencoder finds application in many areas including compression, filtering, and accelerated search. A variational encoder (Kingma & Welling 2013a,b; Rezende et al. 2014) is a probabilistic version of the autoencoder. Rather than mapping the input data to a specific (fixed) approximating vector, it maps the input data to the parameters of a probability distribution, e.g., the mean and variance of a Gaussian. VAEs produce a latent variable $\mathbf{z}$, useful for data generation. We refer to the *prior* for this latent variable by $p(\mathbf{z})$, the observed variable (*input*) by $\mathbf{x}$, and its conditional distribution (*likelihood*) as $p_\theta(\mathbf{x}|\mathbf{z})$. In a Bayesian framework, the relationship between the *input* and the *latent variable* can be fully defined by the *prior*, the *likelihood*, and the *marginal* $p_\theta(\mathbf{x})$ as:

$$p_\theta(\mathbf{x}) = \int_{\mathcal{Z}} p_\theta(\mathbf{x}|\mathbf{z}) p(\mathbf{z}) d\mathbf{z}. \tag{5}$$

It is not not easy to solve Equation (5) as the integration across $\mathbf{z}$ is most often computationally intractable, especially in high dimensions. To tackle this, variational inference (e.g. Jordan et al. 1999) is used to introduce an approximation $q_\phi(\mathbf{z}|\mathbf{x})$ to the posterior $p_\theta(\mathbf{z}|\mathbf{x})$. In addition to maximizing the probability of generating real data, the goal now is also to minimize the difference between the real $p_\theta(\mathbf{z}|\mathbf{x})$ and estimated $q_\phi(\mathbf{z}|\mathbf{x})$ *posteriors*. We state without proof (see Kingma & Welling (2019) for detailed derivation):

$$\begin{aligned} &- \log p_\theta(\mathbf{x}) + D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \,\|p_\theta(\mathbf{z}|\mathbf{x})\right) \\ &= -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] + D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})\right) \end{aligned} \tag{6}$$

The approximating distribution q is chosen to make the right-hand-side of Equation (6) tractable and differentiable. Taking the right-hand-side as the objective to simultaneously minimize both the divergence term on the left-hand-side (making q a good approximation to p) and $-\log p(x)$. This is exactly the loss function that we want to minimize via backpropagation:

$$\begin{aligned} L_{\mathrm{VAE}}(\theta,\phi;\mathbf{x}) &= -\mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x})}\left[\log(p_\theta(\mathbf{x}|\mathbf{z}))\right] \\ &\quad + D_{\mathrm{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x})\|p_\theta(\mathbf{z})\right) \\ -L_{\mathrm{ELBO}} &= L_{\mathrm{REC}}(\theta,\phi;\mathbf{x}) + D_{\mathrm{KL}}(\theta,\phi;\mathbf{x}) \end{aligned} \tag{7}$$

where $\theta^*, \phi^* = \mathrm{argmin}_{\theta,\phi} L_{\mathrm{VAE}}$. Since KL-divergence is non-negative, Equation (7) above can be thought of as the lower bound of $p_\theta(\mathbf{x})$, and is the loss function to minimize. It is commonly called the ELBO, short for *evidence based lower bound*. L_{REC} minimizes the difference between input and encoded samples, while D_{KL} acts as a regularizer (Hoffman & Johnson 2016).

The typical choice for q (that we also make) is an isotropic conditionally Gaussian distribution whose mean and (diagonal) covariance depend on $\mathbf{x}$. The result is that the divergence term has a closed-form expression where the mean and variance are learned, for example by using a neural network. To be able to backpropagate through the first term (the expectation) in the loss function, a reparameterization is introduced. For each sample from $\mathbf{x}$ take one sample of $\mathbf{z}$ from the conditionally Gaussian distribution $q_\phi(\mathbf{z}|\mathbf{x})$. Without loss of generality we can generate an isotropic Gaussian $\mathbf{z}$ by taking a Gaussian source $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ shifting it by the $\mathbf{x}$-dependent mean μ and scaling by the standard deviation $\boldsymbol{\sigma}$ to get $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$, where $\odot$ is the element-wise product. Approximating the first (expectation) term in the objective with a single term using this value for $\mathbf{z}$ allows one to backpropagate gradients through this objective. Note that, in the terminology of autoencoders, the q and p functions play the respective roles of encoder and decoder; $q_\phi(\mathbf{z}|\mathbf{x})$ generate the latent representation from a data point and $p_\theta(\mathbf{x}|\mathbf{z})$ defines a generative model.

The VAE described so far, which we refer to as the 'vanilla' VAE, is not the optimal model for our purposes.This is because our data set $\mathcal{D}_{\mathcal{F}_S,\mathcal{S}_S}$ can contain outliers caused mostly by sensor failures, and sometimes by faulty data processing pipelines. The ELBO for 'vanilla' VAE contains a log-likelihood term (first term in RHS of Equation 6) that will give high values for low-probability samples (Akrami et al. 2019).

We state without proof (see Akrami et al. (2019, 2020) for details) that L_{REC} for a single sample can be re-written as:

$$L^{(i)}_{\mathrm{REC}} = \mathbb{E}_{\mathbf{z}\sim q_\phi(\mathbf{z}|\mathbf{x}^{(i)})}\left[D_{\mathrm{KL}}\left(\hat{p}(\mathbf{X})|p_\theta(\mathbf{X}|\mathbf{Z})\right)\right], \tag{8}$$

where $\hat{p}(\mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N}\delta(\mathbf{X} - \mathbf{x}^{(i)})$ is the empirical distribution of the input matrix $\mathbf{X}$, and N is the number of samples in a mini-batch. We then substitute the KL-divergence with the β-cross entropy (Ghosh & Basu 2016) which is considerably more immune to outliers:

$$L^{(i)}_{\mathrm{REC},\beta} = \mathbb{E}_{\mathbf{Z}\sim q_\phi(\mathbf{Z}|\mathbf{x}^{(i)})}\left[H_\beta(\hat{p}(\mathbf{X}), p_\theta(\mathbf{X}|\mathbf{Z}))\right], \tag{9}$$

[11] https://catboost.ai/

where the β cross-entropy is given by Eguchi & Kato (2010); Roddier (1988); Futami et al. (2018):

$$H_\beta(\hat{p}(\mathbf{X}), p_\theta(\mathbf{X}|\mathbf{Z})) = \\ -\frac{\beta+1}{\beta}\int \hat{p}(\mathbf{X})\left(p_\theta(\mathbf{X}|\mathbf{Z})^\beta - 1\right)d\mathbf{X} + \int p_\theta(\mathbf{X}|\mathbf{Z})^{\beta+1}d\mathbf{X} \quad (10)$$

Here β is a constant close to 0. This makes the total loss function for a given sample:

$$L_\beta\left(\theta, \phi; \mathbf{x}^{(i)}\right) = \mathbb{E}_{\mathbf{Z}\sim q_\phi(\mathbf{Z}|\mathbf{x}^{(i)})}\left[H_\beta(\hat{p}(\mathbf{X}), p_\theta(\mathbf{X}|\mathbf{Z}))\right] \\ + D_{KL}\left(q_\phi\left(\mathbf{Z}|\mathbf{x}^{(i)}\right)\|p_\theta(\mathbf{Z})\right) \quad (11)$$

To draw from the continuous $\mathbf{Z}$, we use an empirical estimate of the expectation, and convert the above into a form of the Stochastic Gradient Variational Bayes (SGVB) cost (Kingma & Welling 2013a) with a single sample $\mathbf{z}^{(j=1)}$ from $\mathbf{Z}$. Next, for each sample we calculate $H_\beta(\hat{p}(\mathbf{X}), p_\theta(\mathbf{X}|\mathbf{z}^{(1)}))$ when $\mathbf{x}^{(i)} \in [0, 1]$. We substitute $\hat{p}(\mathbf{X}) = \delta\left(\mathbf{X} - \mathbf{x}^{(i)}\right)$ and model $p_\theta(\mathbf{X}|\mathbf{z}^{(1)})$ with a mixture of Beta distributions with weight vector ω. That is,

$$p_\theta(\mathbf{X}|\mathbf{z}^{(1)}) = \sum_{k=1}^{k}\omega_k(\mathbf{X}^{p_k-1})(1-\mathbf{X})^{q_k-1} \times \frac{\Gamma(p_k+q_k)}{\Gamma(p_k)\Gamma(q_k)} \quad (12)$$

Using Equations (10) and (12), we obtain:

$$H_\beta(\delta(\mathbf{X}-\mathbf{x}^{(i)}), p_\theta(\mathbf{X}|\mathbf{Z})) = \\ -\frac{\beta+1}{\beta}\left(\sum_{d=1}^{D}\left(\sum_{k=1}^{K}\omega_k(\mathbf{x}_d^{(i)\cdot p_k-1})(1-\mathbf{x}_d^{(i)})^{q_k-1}\Lambda_{d,k}\right)\right) \\ + \sum_{d=1}^{D}\sum_{k=1}^{K}\frac{\left((p_{d,k}-1)^{1+\beta}+1\right)\left((q_{d,k}-1)^{1+\beta}+1\right)}{(p_{d,k}-1)^{1+\beta}+(q_{d,k}-1)^{1+\beta}+2} \quad (13)$$

where $\Lambda_{d,k} = \frac{\Gamma(p_{d,k})+\Gamma(q_{d,k})}{\Gamma(p_{d,k})\Gamma(q_{d,k})}$, D is the number of dimensions in a single sample, and K is the number of components in the mixture. Equations (11) and (13) together give us the total loss across all N samples in a given mini-batch:

$$L_\beta(\theta, \phi; \mathbf{X}) = \frac{1}{N}\sum_{i=1}^{N}\left[H_\beta^{(i)}(\hat{p}(\mathbf{X}), p_\theta(\mathbf{X}|\mathbf{Z})) \right. \\ \left. + D_{KL}\left(q_\phi\left(\mathbf{z}^{(1)}|\mathbf{x}^{(i)}\right)\|p_\theta(\mathbf{z}^{(1)})\right)\right], \quad (14)$$

where the superscript $^{(1)}$ implies a single draw from $\mathbf{z}$ from $\mathbf{Z}$.

The final, robust variational autoencoder architecture is denoted in the left of Figure 4.

4.4 Uncertainty Quantification

Our predictions will be safer for decision making if for each input vector, in addition to the prediction of IQ, we also predict the degree of (un)certainty. This is especially true since we aim to toggle the twelve vents based on our predictions, which is an expensive manoeuvre – a configuration of vents that ends up increasing observed IQ as opposed to decreasing it would require re-observation of the target, when CFHT is already oversubscribed by a factor of ~ 3. For this reason, we predict a probability density function of MPIQ for every input sample, as described in Section 4.1.

Higher error (corresponding to lower model belief or confidence in the estimate) can result from absence of predictive features, error or failure in important sensors, or an input vector that value has drifted from the training distribution. We decompose the sources of predictive uncertainties into two distinct categories: *aleatoric* and *epistemic*. Aleatoric uncertainty captures the uncertainty inherent to the data generating process. To analogize using an everyday object, this is the entropy associated with an independent toss of a fair coin. Epistemic uncertainty, on the other hand captures the uncertainty associated with improper model-fitting. In contrast to its aleatoric counterpart, given a sufficiently large data set epistemic uncertainty can theoretically be reduced to zero[12]. Aleatoric uncertainty is thus sometimes referred to as *irreducible* uncertainty, while epistemic as the *reducible* uncertainty. High aleatoric uncertainty can be indicative of noisy measurements or missing informative features, while high epistemic uncertainty for a prediction could be a pointer to the outlier status of the associated input vector.

The architecture of the MDN (Section 4.1) allows us to predict a PDF of MPIQ for each sample. For each sample and mixture model component, let μ_m, $(\sigma_m)^2$, and α_m respectively denote the mean, variance, and normalized weight (weights for all mixture model components must sum to 1) in the mixture model. We obtain the predicted IQ value as the weighted mean of the individual means:

$$\mu = \sum_{m=1}^{M}\alpha_m\mu_m \quad (15)$$

Aleatoric uncertainty is the weighted average of the mixture model variances, calculated as (Choi et al. 2018):

$$\sigma_{\rm al}^2 = \sum_{m=1}^{M}\alpha_m\sigma_m^2, \quad (16)$$

while epistemic uncertainty is the weighed variance of the mixture model means:

$$\sigma_{\rm epis}^2 = \sum_{m=1}^{M}\alpha_m\mu_m^2 - \mu^2 \quad (17)$$

The total uncertainty is computed by adding (16) and (17) in quadrature.

4.5 Probability Calibration

In Section 4.4 we describe how to derive both aleatoric and epistemic errors. While these variance estimates yield a second-order statistical characterization of the distribution of output errors, they can at times mislead the practitioner into a false sense of overconfidence (Lakshminarayanan et al. 2017; Kuleshov et al. 2018; Zelikman & Healy 2020).

It therefore becomes imperative to *calibrate* our uncertainty estimates to more closely match the true distribution of errors. In other words, we ensure that 68% confidence intervals for MPIQ predictions (derived from the epistemic uncertainty) contain the true MPIQ values $\sim 68\%$ of times. The confidence interval is the range about the point prediction of IQ in which we expect, to some degree of confidence, the true IQ value will lie. For example, if our error were conditionally Gaussian, centered on our point prediction, then we would expect that with about 68.2% probability the true IQ value would lie within $\pm 1\sigma$ of our IQ prediction where σ is the standard deviation of the Gaussian. To accomplish this we reserve some of our data which we use to estimate the distribution of errors – this is

[12] The word "aleatoric" derives from the Latin "aleator" which means "dice player". The word "epistemic" derives from the Greek "episteme" meaning "knowledge" (Gal 2016).

the validation set. Using the inverse cumulative distribution function of this estimated distribution, scaled by the predicted standard deviation and shifted by the predicted mean, allows us to obtain a calibrated estimate of the output realization corresponding to any particular percentile of the distribution. The specific approach we use is the CRUDE method Zelikman & Healy (2020).

However, calibrating the error estimates is not the only thing we care about if, through the calibration process we loose substantial accuracy. For instance, one can increase predicted uncertainties to arbitrarily high values to obtain a perfectly calibrated model; however, this would make these predictions useless for practically any downstream task. Therefore, CRUDE not only calibrates our post-processed predictions, but also ensures that they are *sharp* (Nixon et al. 2019). Sharpness refers to the concentration of the predictions, akin to the inverse of the posterior error variance. The more peaked (the sharper) the predictions are, the better, provided the sharpness does not come at the expense of calibration.

4.6 Performance Metrics

For each input sample $\mathbf{x}$ we derive the predicted IQ, the aleatoric uncertainty, and the epistemic uncertainty, respectively, μ, σ_a, and σ_e, cf. Equations (15), (16), (17). In Section 4.6.1 we compare the median of predicted IQ values against their ground truth values. In Section 4.6.2 we evaluate the quality of the predicted PDF.

4.6.1 Metrics for Deterministic Predictions

We present three measures to quantify the quality of the IQ prediction, Root-mean-square error (RMSE), mean absolute error (MAE), and bias error (BE). Respectively, these three measures are defined as

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mu_i - y_i)^2},$$

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\mu_i - y_i|,$$

$$\mathrm{BE} = \frac{1}{N}\sum_{i=1}^{N}(\mu_i - y_i).$$

In the above definitions, y_i and μ_i are the true and predicted IQ values corresponding to an input sample and N is the number of samples.

4.6.2 Metrics for Probabilistic Predictions

As discussed, for each sample our model yields a prediction tuple $\{\mu, \sigma_a, \sigma_e\}$. We further use σ to denote total uncertainty where $\sigma^2 = \sigma_a^2 + \sigma_e^2$. Considering 68^{th} percentile ("one-sigma") confidence intervals, the lower and upper bounds of the interval are $L_\alpha = \mu - \sigma_e$ and $U_\alpha = \mu + \sigma_e$ where, for this example of a one-sigma confidence interval, CI = 0.682 and $\alpha = 1 - \mathrm{CI} = 1 - 0.682 = 0.318$. The parameter α is the fraction of time the model predicts the true IQ will fall outside the confidence interval. We denote by l_α and u_α the cumulative distribution function of the (assumed Gaussian) PDF respectively evaluated at L_α and U_α, i.e., $l_\alpha = 0.5 - \mathrm{CI}/2 = 0.159$, $u_\alpha = 0.5 + \mathrm{CI}/2 = 0.841$.

We are now ready to introduce our two measures of the quality of our probabilistic predictor: average coverage area (ACE) and interval sharpness (IS). Given N predictions, let the true IQ for one sample be denoted y. We also define an indicator function $\mathbb{1}_\alpha$ that evaluates to 1 if the true IQ of a sample falls within the corresponding predicted confidence intervals, and zero elsewhere:

$$\mathbb{1}_\alpha = \begin{cases} 1 & \text{if} \quad y \in [L_\alpha, U_\alpha] \\ 0 & \text{else} \end{cases}.$$

The average coverage estimator is defined as for all samples:

$$\mathrm{ACE}_\alpha = \frac{1}{N}\sum_{i=1}^{N}\mathbb{1}_\alpha^i - (1-\alpha)$$

and is a measure of the how well the confidence interval captures the realized distribution of predictions. A value of zero tells us that exactly a fraction $1-\alpha$ of the predicted confidence intervals encapsulate the respective true IQs. Generally if ACE_α is small in magnitude then the prediction interval is well matched to the realized distribution of predictions.

While the average coverage area gives us a sense of the match between the predicted and realized distributions, it doesn't give us a sense of the concentration of the error. By letting $\alpha \to 0$ all data points will fall in the bounds and so $\mathrm{ACE}_\alpha \to 0$ too. Therefore we need a second measure of probabilistic prediction. We use interval sharpness/interval score (IS) as this second measure (Gneiting & Raftery 2007; Bracher et al. 2021). Interval sharpness for a single sample is defined as:

$$\mathrm{IS}_\alpha = \begin{cases} \alpha(U_\alpha - L_\alpha) + 2\,[L_\alpha - y] & \text{if } y < L_\alpha, \\ \alpha(U_\alpha - L_\alpha) & \text{if } L_\alpha \le y \le U_\alpha, \\ \alpha(U_\alpha - L_\alpha) + 2\,[y - U_\alpha] & \text{if } y > U_\alpha, \end{cases}$$

We normalize this against similar values for all samples, such that the final value lies between 0 and 1:

$$\mathrm{IS}_{\alpha,\mathrm{norm}} = \frac{\mathrm{IS}_\alpha - \min(\mathbf{IS}_\alpha)}{\max(\mathbf{IS}_\alpha) - \min(\mathbf{IS}_\alpha)}$$

and finally average the normalized values across the samples in the test set:

$$\overline{\mathrm{IS}}_\alpha = \frac{1}{N}\sum_{i=1}^{N}\mathrm{IS}^i_{\alpha,\mathrm{norm}}. \tag{18}$$

To understand Equation (18) we note first that $0 \le \overline{\mathrm{IS}}_\alpha \le 1$ and higher sharpness (less positive) corresponds to more concentration and therefore more useful predictions. The first term, $\alpha(u_\alpha - l_\alpha)$, is a constant, parameterized by α. In our experiments we set $\alpha = 0.318$ corresponding to ± 1 standard deviation. Then a smaller variance will lead to a narrower confidence interval and a smaller IS_α *if* the sample falls within the confidence interval. The sharpness is decreased (IS_α increases) if the prediction y falls outside of the confidence interval, and the penalty applied is proportional to the distance between the ground truth value and the nearest interval limit. Generally a IS_α small in magnitude means the estimates both fall in the confidence interval *and* the confidence interval is narrow.

We calculate ACE and IS for all three uncertainties – aleatoric, epistemic, and total.

4.7 Feature Ranking

One of our goals in this work is to understand the physical mechanisms that yield high and low IQ values so that, in the future, we can actuate the observatory to improve the realized IQ. To accomplish this we need to understand the insights that the ML models decision making processes reveal. To this end, we utilize the methods of integrated Hessians and Shapley values (Janizek et al. 2020; Gilda et al.

2021a,b) for the MDN model. We use an implementation provided by the `pathexplainer` software package which compute feature attributions (or importances). The attributions plot ranks the 119 input features, guiding us on how important each feature is, relative to all other ones, in explaining the predicted MPIQ. These enables us to understand the model's decision making process, and to ascertain that the features deemed important by the model make sense physically.

4.8 Putting It All Together

Training and test sets: For both the MDN (Section 4.1) and the RVAE (Section 4.3) we partition $\mathcal{D}_{\mathcal{F}_S,\mathcal{S}_S}$ into two unequally-sized subsets – a *training super-set* containing 90% of the samples and a *test set* containing the rest. We are following a nested cross-validation scenario. We partition the data sets carefully, to ensure that the distribution of MPIQ values in both the test and training sets reflect the distribution in the original data set. To accomplish this we sort the samples by MPIQ values and, starting from the lowest MPIQ value allocate each sample in a round-robin fashion to one of ten buckets generated. We then iterate this process for the training super-set – again producing a 90-10 split – to respectively produce the final *training set* and the *validation set*. We train the models on the training set and record its predictions on the validation and test sets. The validation set guards against over-fitting – we want our models to learn patterns from the training set, but not to the extent where they fail to generalize to unseen samples. Before making predictions on the test set, we revert the weights of both the MDN and RVAE models to their respective epochs where their respective losses on the validation data set were minimal, as shown in Figure A3b for the MDN. As a quick reminder, a 'prediction' for the MDN is a three-tuple consisting of mean μ, aleatoric uncertainty σ_a, and epistemic uncertainty σ_e for the MPIQ, whereas for the RVAE it is the reconstructed input sample.

Learning rate and optimizer: We use a cyclical learning rate scheduler to vary the learning rate from an initial high to a final low value, in multiple cycles; this has been shown to result in a considerably better convergence than using step-wise or constant learning rate schedules (Smith 2017). To determine these limits for the MDN and the RVAE, we pick arbitrarily high (10^{-1}) and low (10^{-7}) limits, exponentially increase the learning rate from the latter to the former in a mere 20 epochs, and evaluate the behavior of the respective loss functions. For the MDN, we determine that at 10^{-3} and 10^{-6}, the loss begins to plateau, as can be seen from Figure A3a. We thus pick these as the higher and lower limits, respectively, and indicate them by dashed vertical lines. Similarly, from Figure A4a we can see that these limits for the RVAE are 10^{-3} and 10^{-5}. We use the Yogi optimizer (Reddi et al. 2018) for stochastic gradient descent; this optimizer is an improvement over the commonly used Adam (Kingma & Ba 2015), and we find experimentally that it provides faster convergence. We wrap this optimizer in the Stochastic Weight Averaging optimizer (Izmailov et al. 2018) – accessible via the `TensorFlow Addons` library[13] – and average the model weights every 20 epochs, to overlap with the length of a training cycle. The batch size when using both models is 128.

Feature normalization and data augmentation: Finally, we apply strong feature normalization and data augmentation to regularize against over-fitting. Specifically, we use Positional Normalization (Li et al. 2019, PONO) layers to capture both the first and second moments of latent feature vectors, and use Momentum Exchange (Li et al. 2020, MoEx) to mix the moments of one input sample with that of another, to encourage our models to draw out training signal from the moments as well as from the normalized features. In each mini-batch of 128 samples, every feature vector for every sample is added with the feature vector for a randomly picked sample; the probability that this happens is set to 0.5 - this is, half the times, there is no mixing. In case of mixing, the weight assigned to the original sample is picked from a β distribution with both concentration parameters set to 100, while the weight of the randomly picked sample is the difference of this from 1 (so that both weights sum to unity). The same random ordering of samples and the same weights are carried over to the model outputs as well (MPIQ for the MDN, the reconstructed input for the RVAE). This augmentation scheme has shown to produce state-of-the-art results, and our own experiments confirm excellent performance. This can be seen in Figure A3b, where we plot the training and validation losses for one of ten folds; the training loss is significantly higher than the validation loss for a large part of the training process. We insert a PONO layer after each Dense layer in the MDN, and after the penultimate encoding layer in the RVAE. The MoEx layers are inserted before the ultimate Dense layer in the MDN, and the ultimate layer in the RVAE. Each PONO layer is followed by a Group Normalization layer (Wu & He 2018, GN) with a channel size of 16 (see the MDN in Figure 4), except when a MoEx layer directly follows the PONO layer, where the former is followed by a Batch Normalization layer (Ioffe & Szegedy 2015).

Calibration: For the MDN, we implement additional steps to calibrate the predicted MPIQ PDFs. We treat each of the 10 training sets (these are obtained after splitting the respective training *super-sets* into training and validation sets, as explained at the beginning of this section) as a training *super-set*, and the associated validation set as the test set. In other words, we sub-divide the training set into 10 training and validation sets , train on the new training data sets and use the new validation sets as guardrails against over-fitting, and predict MPIQ on the new test sets. After repeating this process a total of 10 times, we now have predictions for the mean and both uncertainties for all samples in the original training set. Finally, we *calibrate* our model's predictions on the original test set by using the predictions on the original training set, by following the method described in Zelikman & Healy (2020). This is the post-processing step discussed in Section 4.5. We repeat this entire process a total of 10 times to cover all samples in $\mathcal{D}_{\mathcal{F}_S,\mathcal{S}_S}$. We illustrate this workflow in Figure A1 in Appendix A, where in the interest of saving space we show only 3 splits instead of 10.

RVAE tuning: For the RVAE, there are a couple of additional considerations. For one, we adopt an annealing methodology to handle the problem of vanishing KL-divergence (Fu et al. 2019). It is known that the KL-divergence loss term in Equation 7 very quickly collapses to 0 if both L_{REC} and L_{KL} are equally weighted. We therefore adopt the methodology suggested by Fu et al. (2019): we modify Equation 7 by multiplying the second term by a weight scalar W_{KL}, and vary this from 0 to 1 in a cyclical fashion, as shown in Figure A4b. Next, there is the requirement to choose an appropriate β in Equation 13. We choose $\beta = 0.005$ based as suggested by Futami et al. (2018), and leave the task of finding an optimal β to future work. Finally, since W_{KL} is annealed with epochs, we need to ensure that our lower and upper learning rates help with convergence for all values of this scalar. From Figure A4a, we see that between learning rates of $10^{(-5)}$ and $10^{(-3)}$, the total loss decreases for all values of W_{KL}.

[13] https://github.com/tensorflow/addons

Overall workflow: Our overall workflow is as follows:

(i) For a given train-test split (out of a total of 10) of $\mathcal{D}_{\mathcal{F}_S,\mathcal{S}_S}$, we use the training set with the MDN, record predictions on the test set, and calibrate them using the methodology described above. We save the weights of the MDN at the epoch of minimum validation loss – this is shown by the dashed vertical line in Figure A3b, and for the specific split shown, occurs at epoch 38.

(ii) Next, we train the RVAE using the same training set. Similar to the process with the MDN, we revert the model weights back to the epoch of minimum loss, and make predictions on the test set. We gather for the training, validation, and test sets the total loss -L_{ELBO}, reconstruction loss L_{REC}, and the KL-divergence loss L_{KL}. These are plotted in Figures A4c and A4d. We save the 95th percentile of -$L_{ELBO,Train}$ as the L2; this is our cut-off between ID and OoD samples.

(iii) Next, we create a small data set of only those samples from the test set where all twelve vents are open. While our goal is to hypothesize the gains in seeing/MPIQ we could have gotten had the vents been in their optimal configuration instead of in the all-open configuration, we believe it is important to be conservative in our estimates. Thus we select only those samples for further processing where we are confident that there were no mechanical malfunctions, high wind conditions, or other system errors that could have prevented the telescope operator from opening all vents.

(iv) As a first filter, we select only those samples for which $L_{ELBO,Test}$ < L2, with the intention of filtering out samples for which we are not extremely confident about the ID characteristic.

(v) From this newly created test set, we further only select those samples where our MDN from Step (i) predicts that the true MPIQ is covered by 68^{th} percent spread about the median in the predicted MPIQ PDF. This is again enacted in the interest of obtaining conservative predictions downstream.

(vi) From the filtered test set in Step (v), we create a permutated data set by toggling all twelve vents ON (==1) and OFF (==0). For a total of 12 vents, this results in 4095 new samples for each input sample, where the remaining 107 features remain unchanged. The 4096th sample is the input test sample itself, since its vents are already in the all-open configuration. For each of the these 4095 samples, we again apply the same filter as in Step (iv) – filtering out those vent configurations which, given the training set, are OoD.

(vii) Finally, we obtain MPIQ predictions using the MDN for all samples in the permutated data set, created by collation ID permutations for all selected test samples.

Identifying predictable vent variations & separating in-distribution from out-of-distribution samples: In Figure 8a, we demonstrate our methodology for separating ID samples from OoD ones. As should be expected, most test samples are ID, as are 95% training samples (by definition). A striking yet expected result is that only a very small sample of possible permutations are ID. The reason for this becomes clear from Figure 11c, where we plot histograms of the different vent configurations in the training set – 0 on the x-axis corresponds to the all-open configuration, while 1 to all-closed. The vast majority of samples, ~ 80%, have all vents closed, while ~ 20% have either all vents or most vents open. Thus the vast majority of samples in the permutated dataset, where the twelve vents can take arbitrary configurations – say half open and half closed, corresponding to a Hamming distance (x-axis in Figure 8a) of 0.5 – are those that the RVAE has not seen before, and thus classifies as OoD.

Process illustration: Finally, we illustrate the workflow delineated in Steps (ii) through (vii) above in Figure A2.

5 RESULTS

In Section 5.1 we present results on using our model to predict the image quality given the current environmental and dome operating conditions. In Section 5.2 we discuss how we might better operate the dome to improve IQ. In particular we investigate the potential improvement that could result from smart actuation of the configuration of the dome vents. In Section 5.3 we present results on the relative contribution of different features to the predicted mean MPIQ. Through these results we verify observations by earlier groups and we start to understand better what information our models use in its inference process.

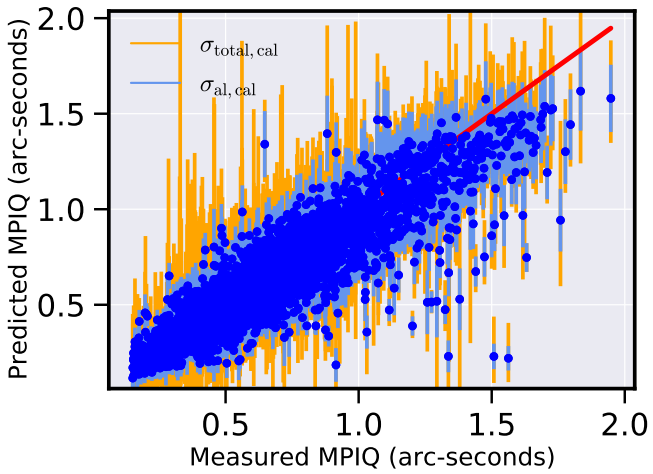
5.1 Predicting image quality

In Figure 5 we present our main results on the accuracy of probabilistic predictions of MPIQ using the MDN. In Figure 6 we present comparative results for the graient-boosted tree model. Table 5 tabulates summary results. We describe each set of results in turn.

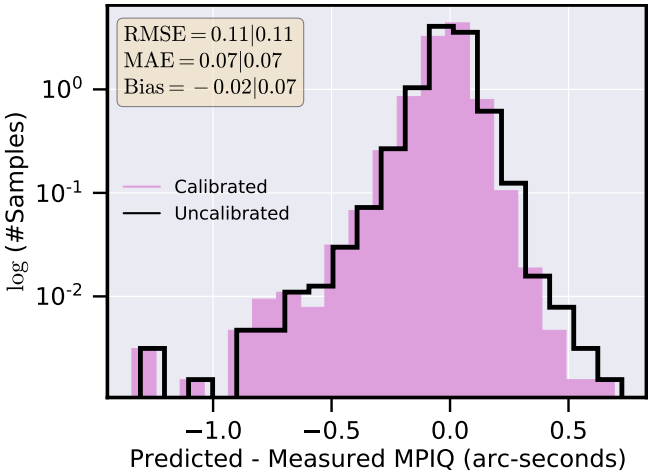
Figure 5a quantifies the accuracy of our predictions. The horizontal axis displays measured (a.k.a. nominal) MPIQ, while the y-axis displays predicted MPIQ. The units of both are arc-seconds ($''$). Perfect prediction is represented by the red 45° line. True MPIQ varies from a bit below $0.5''$ to just over $2''$. The blue dots depict the point-predictions (the medians of the output PDFs). The light blue bars plot the estimated aleatoric uncertainties (σ_a) of the point predictions. These are superimposed on the total uncertainty, the differences are the epistemic uncertainties (σ_e), visible in orange. As is tabulated in Table 5, the mean absolute error (MAE) between the true MPIQ values and the medians of our calibrated predictions is ~ $0.07''$.

Figure 5b help us understand the improvement due to calibration. We plot the histograms of the differences between the calibrated predictions and the true MPIQ values, and between the uncalibrated predictions and the true MPIQ values. These histograms are respectively plotted in pink and black. We use three metrics (cf., Section 4.6.1) to quantify the improvement resulting from calibration: root mean squared error (RMSE), mean absolute error (MAE), and bias error (BE). The values in the first row are for uncalibrated medians while those in the second row are for calibrated models. We remind the reader that the calibration using CRUDE (Zelikman & Healy 2020) is enacted only for the epistemic uncertainties, σ_{epis}, which we observe is significantly decreased for the calibrated model.

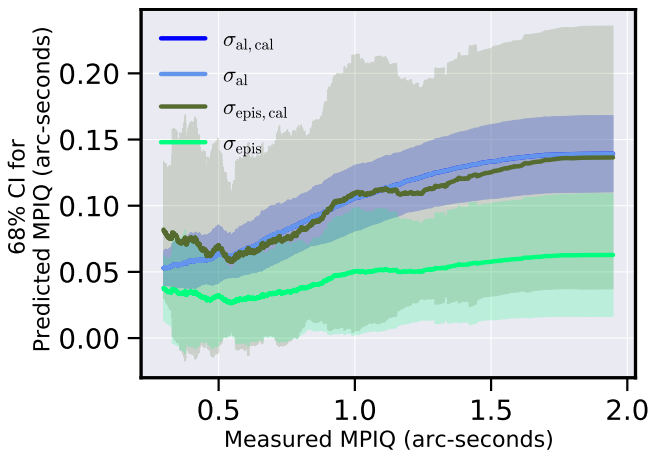
Figures 5c and 5d show smoothed averages of the aleatoric, epistemic, and total uncertainties, for both calibrated and uncalibrated models. We highlight a few important aspects. First, as expected, σ_{al} is unchanged by calibration since we do not calibrate aleatoric uncertainty. (The light-blue and dark-blue plots coincide so we don't see both.) Second, as a function of increasing MPIQ, σ_{al} (and the identical $\sigma_{\text{al,cal}}$ curve) start from a low value, decrease slightly, and then increases almost 3×. The initial dip can be attributed to the high density of data points near the mode of the MPIQ distribution at ~ $0.7''$. The increase at higher MPIQ is likely due to the decreasing density of data points (see the red curve in the left sub-figure of Figure 2). As the model has access to fewer and fewer points it becomes challenging to learn latent representations discriminative enough to be able make good predictions. Hence, the aleatoric error increases. Third, comparing σ_{epis} to $\sigma_{\text{epis,cal}}$ we observe that calibration increases epistemic uncertainty. This justifies our suspicion that the probabilistic MPIQ predictions are over-confident, and that our decision to calibrate them *post-hoc* was



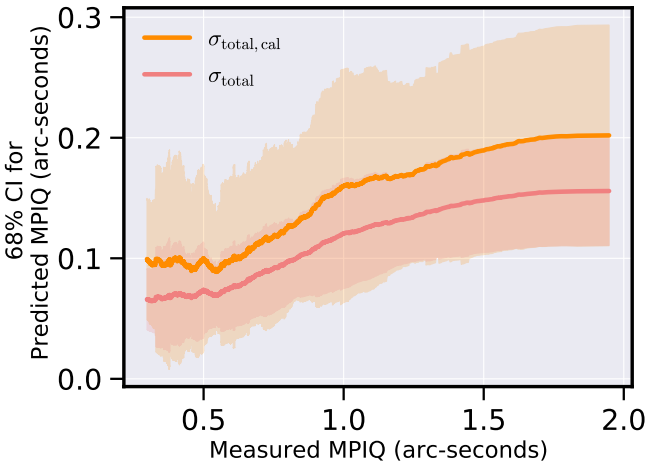
(a) Predicted versus measured MPIQ. For the former we show the calibrated median, and upper and lower quantiles for calibrated uncertainties.



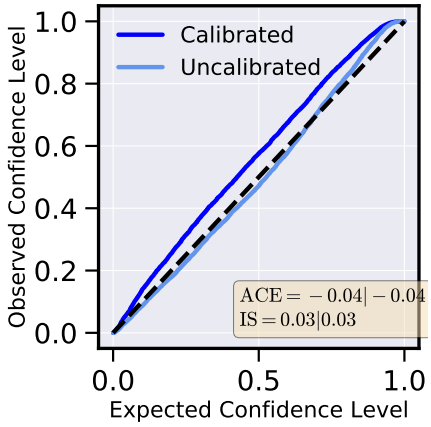
(b) Histogram of prediction errors, along (in inset box) with three metrics that compare performance with and without calibration.



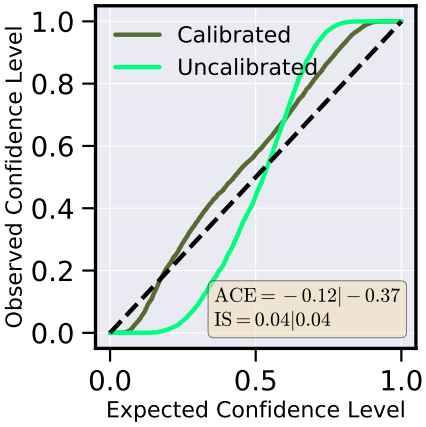
(c) 68% spread in calibrated and uncalibrated aleatoric and epistemic uncertainties in predicted MPIQ, plotted as a function of measured MPIQ.



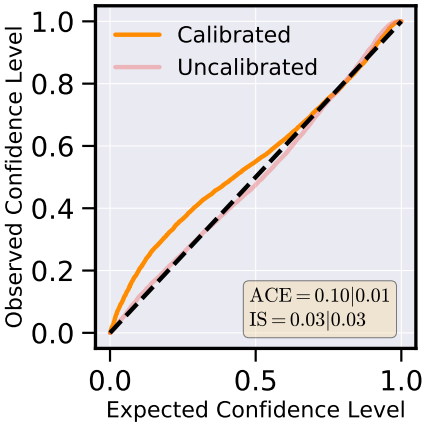
(d) 68% spread in total uncertainties in predicted MPIQ, plotted as a function of measured MPIQ.



(e) Calibration curves for aleatoric uncertainty.

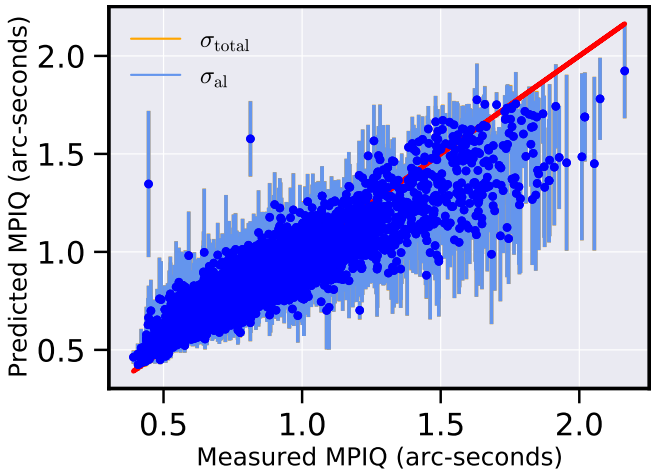


(f) Calibration curves for epistemic uncertainty.

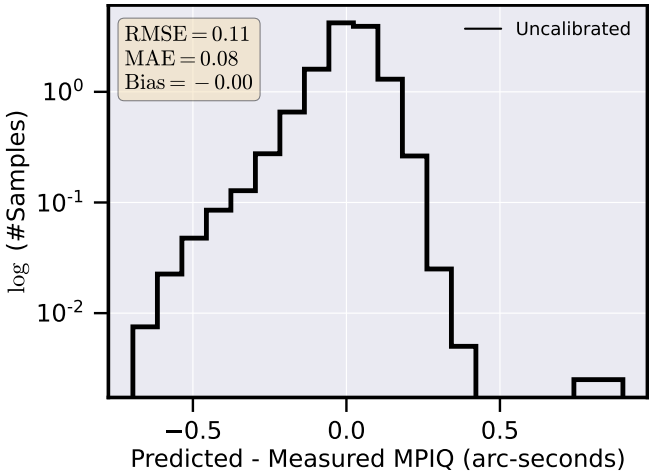


(g) Calibration curves for total uncertainty.

Figure 5. Predictions using the Mixture Density Network. $\sigma \implies 0.5 \times (84^{\text{th}} - 16^{\text{th}})$ quantiles. In **(a)** we plot the predicted MPIQ, characterized by the 16^{th}, 50^{th}, and 84^{thn} percentiles of their respective calibrated PDFs. These are plotted versus measured (i.e., true) MPIQs. In **(b)** we subtract the ground-truth MPIQ from the 50^{th} percentile predictions, from both the raw uncalibrated, and the calibrated PDFs, and plot their histograms. We also quantify the quality of both calibrated and uncalibrated predictions using root mean squared error (RMSE), mean absolute error (MAE), and bias error (BE); in the inset box read calibrated on left and uncalibrated on right. Calibration results in better BE. In **(c)** and **(d)**, we plot the smoothed mean and standard deviations of the aleatoric, epistemic, and total uncertainties as a function of the measured MPIQs. All three uncertainties increase when σ_{epis} is calibrated. Uncertainty is highest near the low-end and high-end MPIQ values; in these regimes we have the least number of observations. Uncertainty is lowest near the mode of the histogram of measured MPIQs where data is plentiful (cf., the histograms in Figure 2). Finally, in **(e)**, **(f)** and **(g)** we visualize the benefits of calibrating σ_{epis}. The ideal is the 1:1 line; closer is better. In the inset box the interval sharpness (IS) and average calibration error (ACE) metrics, with and without calibration, are provided.



(a) Predicted versus measured MPIQ. We do not calibrate the predictions from the boosted tree model.



(b) Histogram of prediction errors. In the inset box we provide three metrics that compare performance with and without calibration.

Figure 6. Prediction results using gradient boosted decision trees (GBDTs).

Table 5. Comparative performance of methods, across all four data sets (cf. Section 3), and all six metrics (cf. Section 4.6). The top row shows uncalibrated results for the MDN and the GBDT models. To ease direct comparison we present results as an ordered tuple (MDN, GBDT). The bottom row shows calibrated results for the MDN. We do not provide calibrated results for the GBDT. For each metric the performance of the best-performing model is highlighted in bold; in all cases the MDN performs at least as well as the GBDT.

Metric	RMSE	MAE	BE	ACE_{al}	ACE_{epis}	$\text{ACE}_{\text{total}}$	IS_{al}	IS_{epis}	IS_{total}
Uncalibrated	(**0.11**, **0.11**)	(**0.07**, 0.08)	(**0.00**, **0.00**)	(**-0.01**, -0.09)	(**-0.31**, -0.59)	(**0.04**, -0.08)	(**0.03**, 0.06)	(**0.04**, 0.08)	(**0.03**, 0.06)
Calibrated	0.11	0.07	0.03	-0.02	-0.10	0.11	0.02	0.04	0.03

sensible. Fourth, σ_{epis} and $\sigma_{\text{epis,cal}}$ follow the same pattern as the aleatoric uncertainty; they initially dip to a minimum and then rise with increasing true MPIQ. That said, relative to their starting values, they dip down to lower levels, and rise asymptotically to about 1.25× their respective starting levels. Since epistemic uncertainty quantifies the degree to which a sample is out-of-distribution (OoD), these curves imply that, compared to the samples near the median MPIQ of 0.7″, samples at both the low and high ends of the measured MPIQ distribution are slightly OoD. (We do note that using predicted epistemic uncertainties is not an reliable way to filter out OoD samples, as expounded upon in Section 5.2 and Figure 7). We believe that both σ_{al} and σ_{epis} can be reduced by weighing the loss function for the MDN so that samples with poorer predictions are given more attention by the network. Another strategy would be to over- and under-sample data points near the ends and the mode of the MPIQ distribution, respectively. This will make the curve be less peaked. By attacking the class-imbalance problem at both the algorithm- and data-level, we expect to de-bias our predictions.

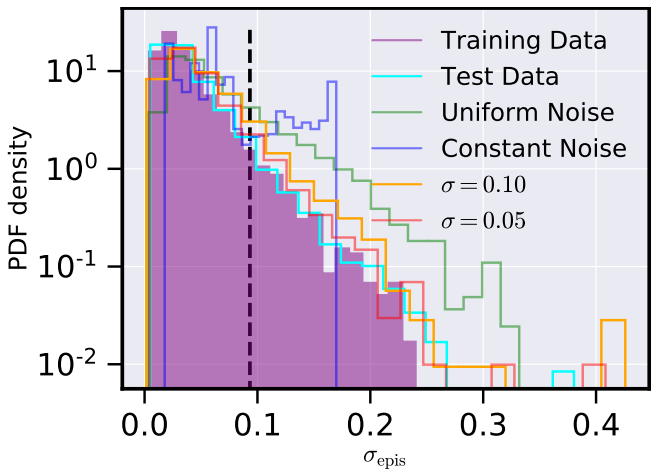
Finally in Figures 5e, 5f and 5g we demonstrate the effect of probability calibration on the three uncertainties. The x- and y-axes respectively quantify the expected and observed confidence levels. If we sample the 50% CI spread around the median of the predicted MPIQ PDF from the MDN, 50% of samples should have their measured MPIQ values be covered by the predicted intervals. Hence the black dashed 1:1 line in all three plots is the ideal calibration plot. In the inserts, we also quantify the difference that calibration makes via the ACE and IS metrics, defined in Section 4.6.2. The values to the left of the vertical bar ('|') in the wheat-colored inserts are for calibrated results, while those to the right for uncalibrated results. Since we only calibrate σ_{epis}, only Figure 5f shows an improvement. This comes at the cost of poorer post-calibration results for σ_{al} and σ_{total}.

In Figure 6 we show comparative plots for predictions made using our gradient-boosted tree models. This model is described in Section 4.2. Comparing Figure 6a to Figure 5a, we see that GBDT models significantly underestimate σ_{epis}. Calibrating the GBDT model using CRUDE does not result in substantial improvement. This is why we use the MDN as our workhorse for MPIQ predictions.

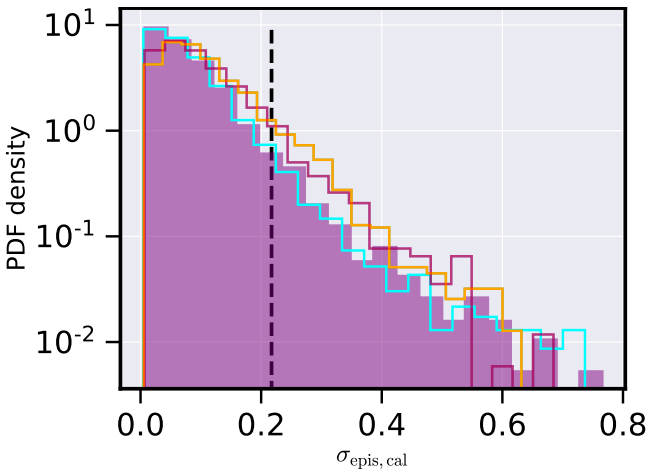
In Table 5 we collate the results on the five metrics, for both calibrated and uncalibrated predictions from the MDN. We compare these predictions from those from the boosted-tree GBDT model. These results demonstrate that the MDN outperforms the GBDT, again supporting the choice to use it as the workhorse model for MPIQ prediction.

5.2 Actuating dome parameters to improve IQ

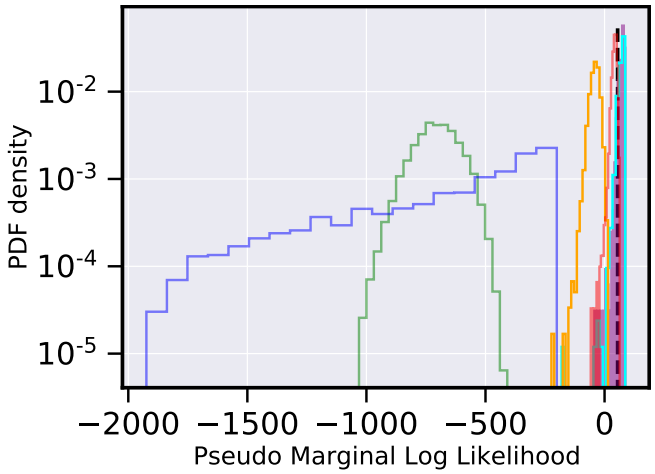
In addition to predicting MPIQ, one of our driving motivations is to learn how to actuate observatory operating parameters to improve MPIQ. One set of easily actuatable parameters is the dome vents. Indeed, as mentioned in the discussion of related work in Section 2, fluid flow models were developed in the vent design process to predict the effect on MPIQ of various vent configurations. These models predicted that the optimal MPIQ is achievable with intermediate vent configurations, where the 12 vents are neither all-open nor all-closed (Baril et al. 2012). In contrast, in most usage to date vents have been configured either to the all-open or to the all-closed setting. We therefore explore what our MDN model predicts – how much improvement a modified vent configuration might have on MPIQ reduction. We note that we must be cautious when



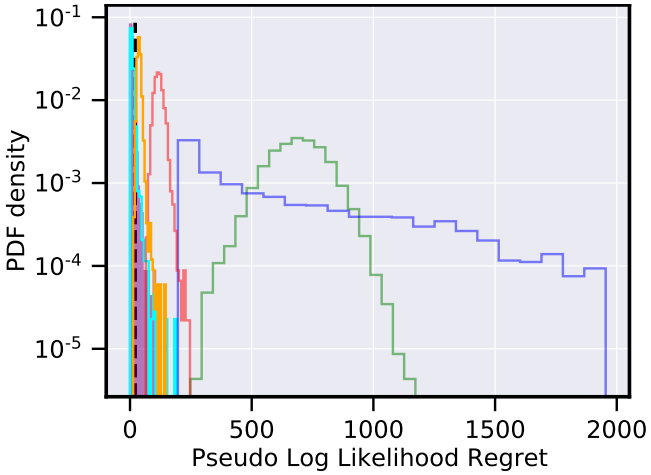
(a) Histograms of *uncalibrated epistemic* uncertainty from the MDN, for various data sets. The vertical line is the 95$^{\text{th}}$ percentile for the training set.



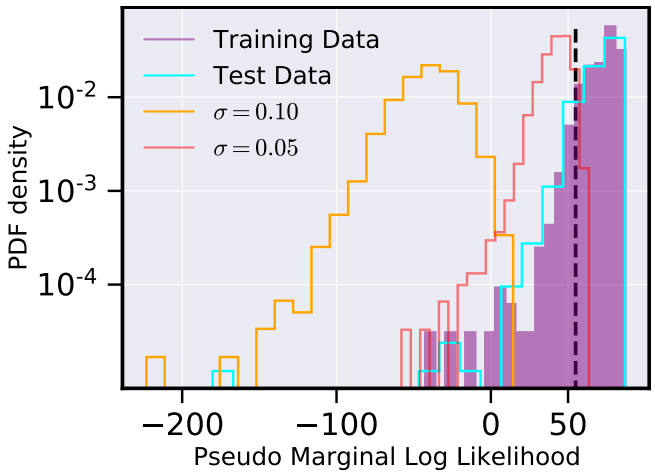
(b) Histograms of *calibrated epistemic* uncertainty from the MDN, for various data sets. The vertical line is the 95$^{\text{th}}$ percentile for the training set.



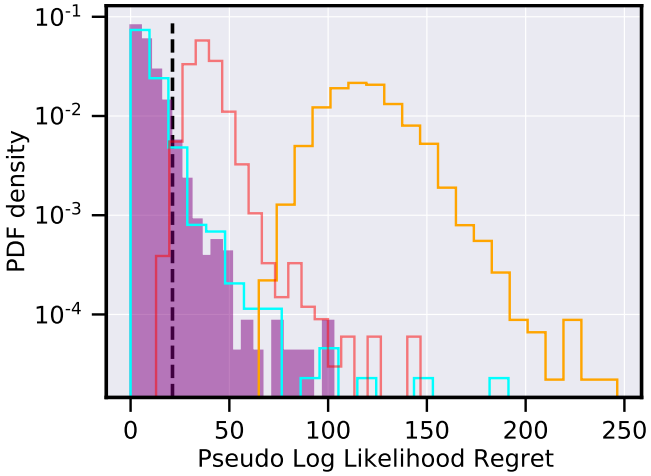
(c) Histograms of the pseudo marginal log likelihood (-L_{ELBO}) from the RVAE. The vertical line is the 95$^{\text{th}}$ percentile for the training set.



(d) Histograms of log likelihood regret for the same data sets, from the RVAE. The vertical line is the 95$^{\text{th}}$ percentile for the training set.

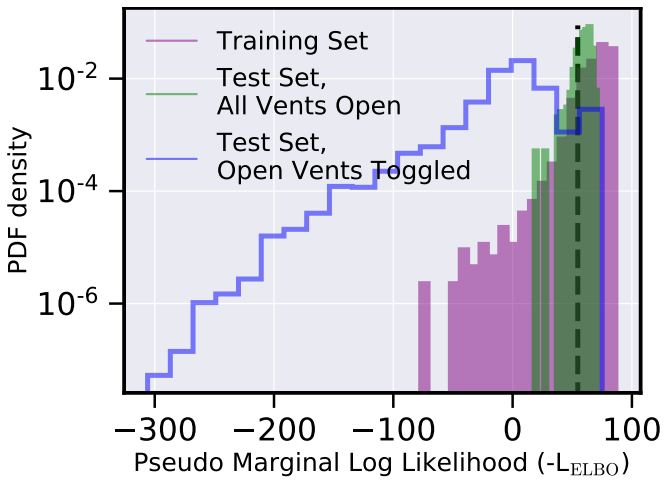


(e) Zoom-in of subfigure **(c)**, focusing on the hard OoD data sets.

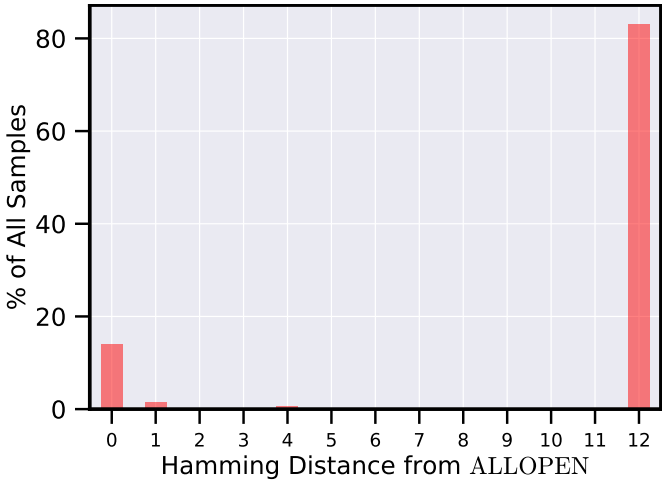


(f) Zoom-in of subfigure **(d)**, focusing on the hard OoD data sets.

Figure 7. We visualize the discriminative ability of various metrics to separate Out-of-Distribution (OoD) samples from In-Distribution (ID) samples. Perhaps surprisingly, we observe that epistemic uncertainties, whether calibrated or uncalibrated, are poor metrics for OoD detection. This is our motivation for the RVAE, which directly captures the likelihood of the training data-generating distribution. **(c)** and **(e)** show that log-likelihood is an excellent, if not perfect, discriminator that can separate ID training and test data from even slightly OoD synthetic data created by adding Gaussian noise of the indicated standard deviation (σ) to normalized training data. In **(d)** and **(f)** we calculate the log-likelihood regret (Xiao et al. 2020), as explained in Section 4.3. Comparing **(e)** and **(f)**, we see that regret is a slightly better OoD detector. In these plots the training data histogram is more concentrated, and the modes of the histograms of the two noisy data sets are farther away from the mode of the histogram of the training data set.



(a) Histograms of the pseudo marginal log likelihood (-L_{ELBO}) for three data sets based on one of ten training folds: the training set, the subset of the test set where all vents are open, and the permutated data set constructed by toggling each of the twelve vents for each sample in the test subset. The dashed line indicates the 95th percentile for -L_{ELBO} of the training data set.



(b) For the training set (purple curve in **(a)**), we visualize the percentage of observed vent configurations by plotting their Hamming distance from the all-open configuration. With twelve vents, each of which can be either open or closed, we can only have a total of thirteen Hamming distances, with 0 denoting all-open and 1 denoting all-closed.

Figure 8. Identifying valid, ID samples from the corpus of all toggled vent configurations. From samples in the test set with all twelve vents open, we select only those about which we can make confident predictions of MPIQ using our MDN. Only a very small subset of the blue curve is ID. This makes sense, since the training set consists mostly of samples where the vents are almost all-open or all-closed. Hence, most samples in the toggled data set with other vent configurations are classified as OoD by our RVAE.

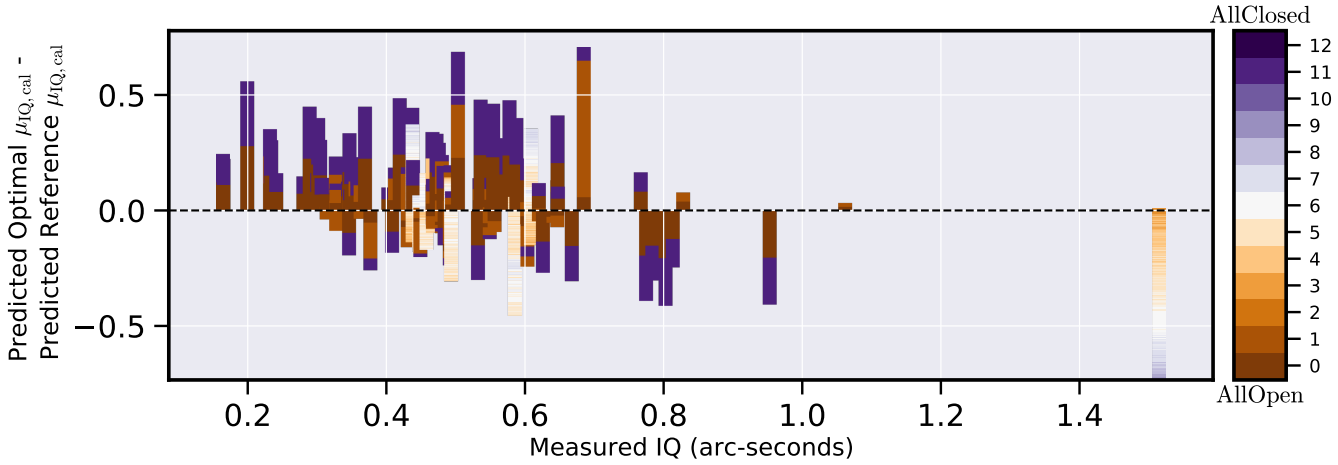
pursuing this exercise as some vent configurations are not within the training sample. As we describe in Section 4 and Figure 7c, we use the pseudo marginal log likelihood, -L_{ELBO}, from the RVAE model as a filter to discriminate in-distribution samples from out-of-distribution ones. In Figure 7 we justify our choice to use this metric to detect distribution shift.

In Figure 7a the pink and cyan curves are the histograms for σ_{epis} for the training and test sets for one of ten folds. To simulate out-of-distribution data, we synthesize four data sets. The uniform noise data set, depicted in green, is generated by drawing $50,000 \times 119$ samples, independently, from the uniform distribution between 0 and 1. The constant noise data set, depicted in blue, is generated by drawing $50,000 \times 1$ samples, independently, from the uniform distribution between 0 and 1, and copying this over 119 times. The orange and red curves are noisy versions of the training data set, where we add Gaussian noise with $\mu = 0$ and $\sigma = 0.10$ and 0.05, respectively. Since we do not train the MDN with noisy versions of the training data (we use the MoEx data augmentation method only, as described in Section 4), the uncertainty in predicting MPIQ as a result of noisy versions of training data is classified as epistemic and not aleatoric. The dashed vertical black line marks the 95th percentile value for $\sigma_{epis,train}$ – we classify all values to its right as out-of-distribution. We plot the density in log scale to better capture different ranges. Figure 7b is the same as Figure 7a, except it plots histograms for calibrated epistemic uncertainty. In both figures, it is apparent that epistemic uncertainty, whether calibrated or uncalibrated, is a poor detector of a distribution shift. Distribution shift identification using discriminative models such as the MDN is an area of active research, and we relegate further exploration of this limitation to future work. In this paper, we instead use the RVAE as a proxy for our data distribution, and justify our decision in Figures 7c, 7d, 7e, and 7f.

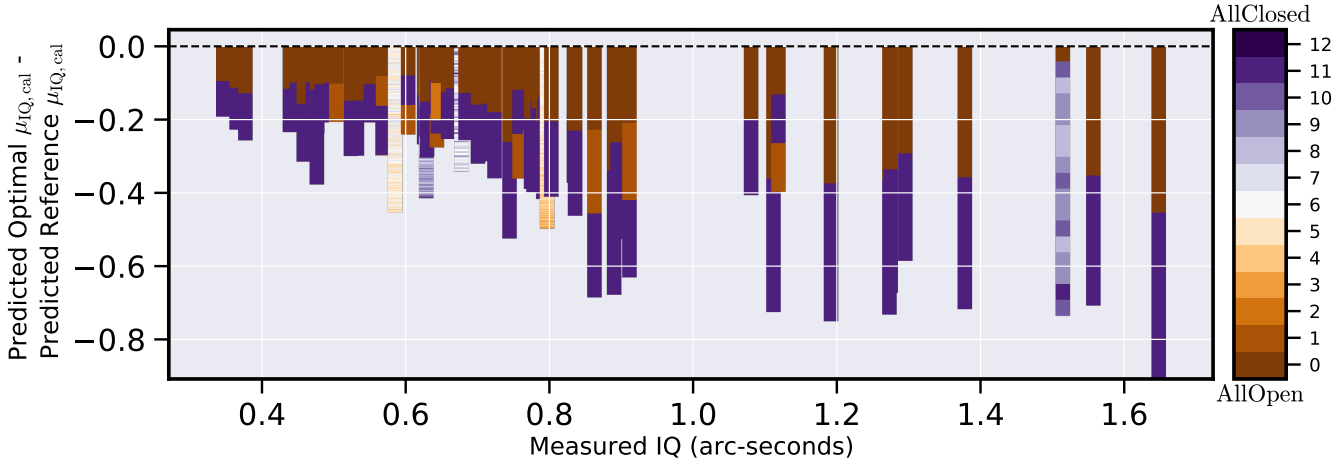
In Figures 7c and 7e, we plot the L_{ELBO} for the same data sets described above; the latter figures focuses on the "hard" cases of noisy training data. The samples to the left are out-of-distribution. It is immediately clear that the marginal likelihood is a much better distribution shift detector than is the epistemic uncertainty. This follows because the uniform and constant noise data sets are very easily separable, and the supports of the two Gaussian datasets are both also almost completely to the left of the vertical line. In Figures 7d and 7f, instead of using L_{ELBO} as the discriminative metric, we instead use the pseudo log likelihood regret. Proposed in Xiao et al. (2020), likelihood regret for a sample is derived by passing it through the trained RVAE and caching L_{ELBO}. We then fix the weights of the decoder and train the encoder to minimize L_{ELBO} for that single sample. The difference between $L_{ELBO,sample}$ and L_{ELBO} is the pseudo log likelihood regret. In Xiao et al. (2020) this is shown to be a better detector of out-of-distribution samples than L_{ELBO}. By definition, regret is always non-negative. Comparing Figures 7e and 7f, we verify that regret is indeed a better separator – the pink curve is less spread out, and the modes of both the red and orange curves are farther away from the black line. Given these results, it is natural to question our design choice of using L_{ELBO} as the metric we used to identify out-of-distribution samples. We use L_{ELBO} rather than regret because while pseudo log likelihood regret is more robust, it also takes about 50 times longer to calculate than L_{ELBO}. This is because the calculation of regret requires retraining of the encoder, once per input sample. Even if we divide our GPU[14] into multiple virtual cores for parallel processing, it took about two hours to calculate regret for 4096 samples ($4096 = 2^12$ is the number of permuted vent configurations possible for each input sample). For these computational reasons we use L_{ELBO}. Depending on computing resources, in the future we may move to distributed computing framework to make the use of regret practical.

We now use L_{ELBO} to identify identify the vent settings that are not too "out-of-distribution" for which our model will be able to

[14] NVIDIA Titan RTX 24Gb



(a) Predicted change in MPIQ corresponding to each possible vent configuration for 100 randomly selected samples, actuating vent configurations across all in-distribution settings for each. Lower is better. The general trend is that the higher the measured IQ the more room there is for improvement.



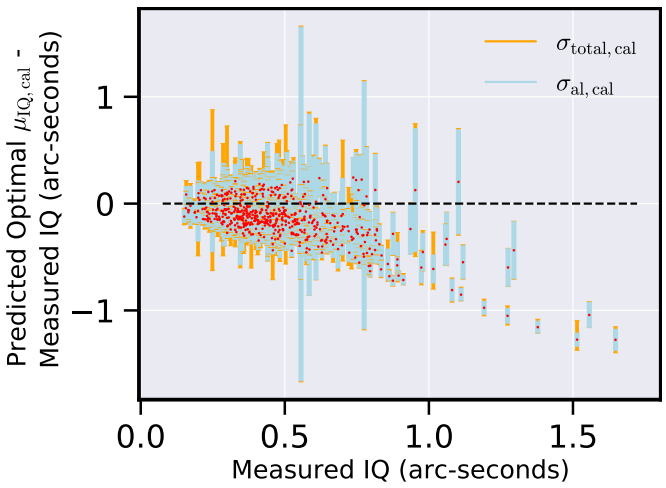
(b) A similar plot to **(a)** except that now we restrict ourselves to a robust sample, as predicted by total uncertainty. Of the roughly 600 samples, according to this measure only 62 are robust. For these we see that the predicted change is an improvement (a decrease) in IQ. Note that this subset of 62 samples need not overlap significantly or at all with those samples in **(a)**.

Figure 9. Visualizing the predicted effect on IQ of optimizing the vent configuration.

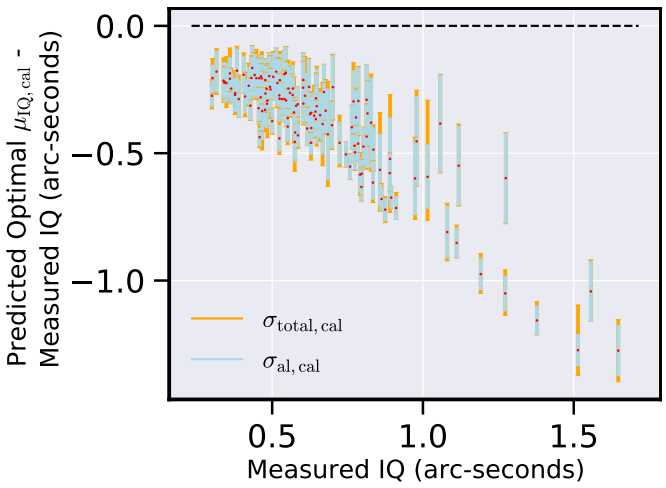
make reliable MPIQ predictions. As we will develop in the following, these robust predictions indicate that substantial MPIQ improvement is possible by optimizing the vent configuration. In future work we plan to extend our dataset to reduce the set of out-of-distribution vent configurations, thereby enabling a wider range of reliable predictions, and helping us to realize even greater MPIQ improvements.

Figure 8 demonstrates results from the process we use to identify, among all possible vent configurations, those for which we can make reliable MPIQ predictions. By this process we filter out those data records that are OoD. (The workflow that led to these results was described towards the end of Section 4.8 and is illustrated in Figure A2.) Is is this restricted, or "filtered", set of vent configurations that we use to assess the possible improvement. In Figure 8a we show results for one of the ten splits of $\mathcal{D}_{\mathcal{F}_S,\mathcal{S}_S}$ and predict the MPIQ that would results for all possible vent settings. On average, each test split of $\sim$ 6000 data records results in $\sim$ 600 samples. For each of these 6000 samples on average only about 4 other vent configurations (out of a possible $2^{12} - 1 = 4095$) are not OoD given the training distribution. For each of these vent configurations we use the MDN to predict the MPIQ three-tuple (μ, σ_{al}, and σ_{epis}). We compare these predictions to the MPIQ three-tuple predictions for the respective base samples (with all vents open). The results of this exercise, which we will discuss next, are presented in Figures 9, 10 and 11.

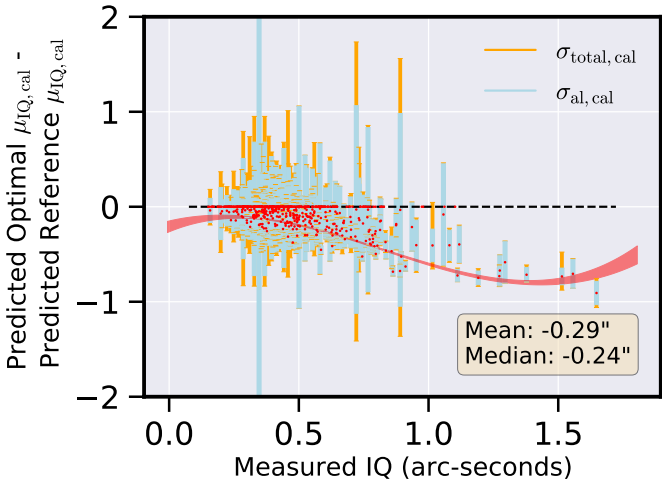
In Figure 9a we plot bar-charts for the change in MPIQ with optimal vent configurations, with respect to the predicted (calibrated) mean MPIQs for the respective reference test samples with all vents open. For each of the 10 train-test splits of $\mathcal{D}_{\mathcal{F}_S,\mathcal{S}_S}$ we recall that there are $\sim$ 600 viable all-open data samples. Of these viable samples we randomly sub-sample 100 and, for each of these, predict their mean MPIQ for the all-open configuration using our MDN. We then calibrate these predictions using the CRUDE method, as described in Section 4.5. We note this prediction will be somewhat different from the measured MPIQ associated with these data records. Finally, we make predictions for each of the ID vent configurations for all 100 samples (roughly $100 \times 4 = 400$) and subtract each of these MPIQ predictions from the predicated calibrated median MPIQs for the all-open configuration. In Figure 9a we plot these difference against the measured MPIQ values. Values above dashed zero-level imply a worsened (predicted) MPIQ in comparison to the baseline of keeping all twelve vents open. Values below



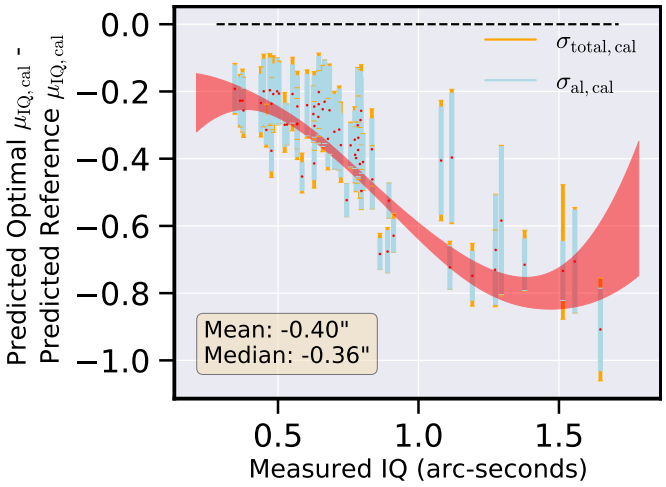
(a) Improvement in MPIQ predicted by the MDN for the optimal vent configurations. We plot with respect to the measured MPIQ. Negative is better.



(b) Same as **(a)**, but only for those samples where the 84$^{\mathrm{th}}$ percentile of predicted optimal MPIQ ≤ measured MPIQ.



(c) Same as **(a)**, with the change that here the improvement is calculated with respect to the predicted MPIQ for the *all-open* samples.



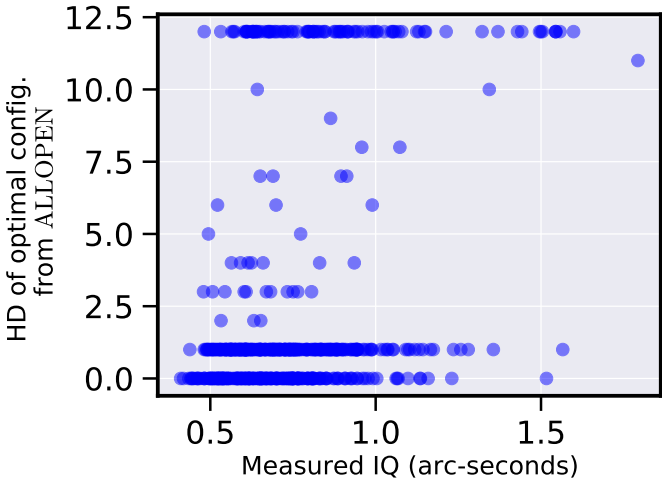
(d) Same as **(c)**, but for only those samples where the 84$^{\mathrm{th}}$ percentile of predicted optimal MPIQ ≤ 50$^{\mathrm{th}}$ percentile of predicted reference MPIQ.

Figure 10. We visualize the gains in terms of improved MPIQ prediction that can be achieved by our proposed methodology of toggling the twelve vents open and close individually as a function of environmental and observatory characteristics. The baseline configuration is all-open. After restricting ourselves to a subset of in-distribution "togglings", in **(a)** and **(b)** we plot the improvement over the measured MPIQ values, whereas in **(c)** and **(d)** we plot the improvement over MPIQ values predicted for the same samples in *all-open* configuration. In **(b)** and **(d)**, we sub-sample the data points from **(a)** and **(c)** respectively, only presenting those samples for which we are quite confident in our estimates. In **(c)**, the several $y = 0$ red dots in the left-half of the plot signify that for those samples, the *all-open* vent configuration is in fact the optimal vent configuration. Finally, we present third-order polynomial fits in **(c)** and **(d)**, and estimate total gains achievable using our predicted, optimal vent configurations, using weighted mean and weighted median metrics. These fits are used to calculate average expected reduction in observing times to achieve a fixed signal-to-noise ratio (SNR).

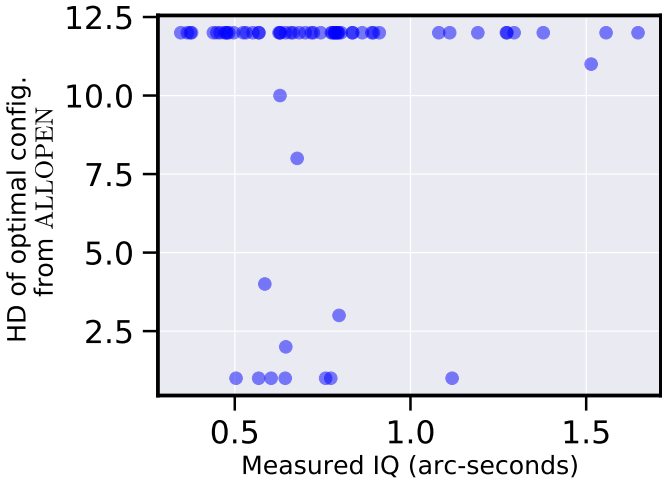
the dashed zero-level suggest that another ID vent configuration will likely result in reduced seeing. Note that we predict the difference in *predicted* seeing levels as these are the levels the model would predict were it *not* to have a measurement of MPIQ for the baseline all-open configuration. While in our data set we *do* have the baseline MPIQ, in real-time operation that baseline MPIQ value would not be available prior to the observation when the observer would be using the model to decide how to actuate the vent configurations.

To better understand how to read the vent configurations that lead to an improvement in predicted MPIQ, we have color-coded Figure 9a. Bars that are dark purple correspond to the all-close configuration; dark brown to all-open. The color gradient corresponds to Hamming distances of the configuration vectors from all-open. As one would suppose, the model predicts all-closed to be a worse setting more often than not. This is in keeping with the original motivation for installing the vents, discussed in Section 2. By and large, opening vents improves MPIQ by allowing air currents built up inside the dome to flush. All-close is the same as having no vents, thus represents the scenario that was meant to improve upon by installing the vents. As we consider higher measured MPIQs (moving from left to right on the x-axis), we see that the optimal configurations tend to be closer to all-close. This is also in line with intuition previously developed at CFHT. Higher measured MPIQs are typically obtained in high wind speed scenarios, where it makes sense to close the vents.

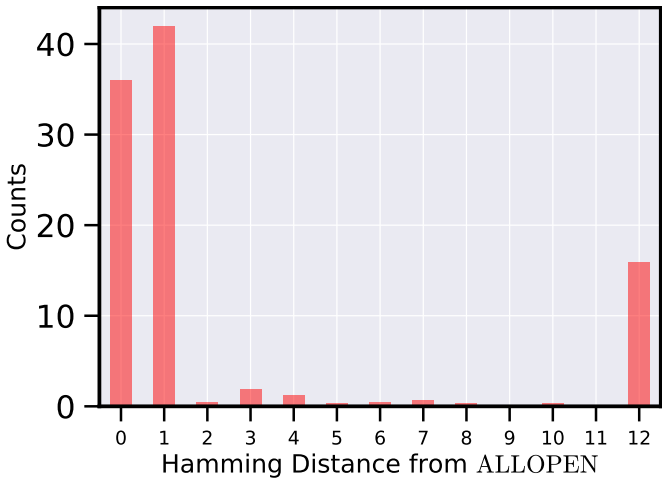
Figure 9b consists of *only* those test samples from the same ∼ 600 that were used to draw from in Figure 9a, where the predicted, calibrated upper 84$^{\mathrm{th}}$ quantiles of the total uncertainties ($\mu_{\mathrm{cal}} + \sigma_{\mathrm{total,cal}}$) are ≤ the predicted calibrated medians for the respective test samples with all vents open. Our idea here is to create



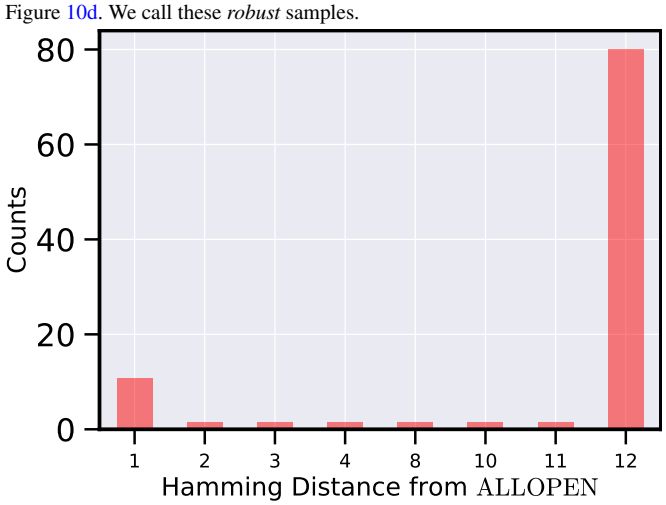
(a) Hamming distance of optimal vent configurations from the all-open configuration. We plot versus measured MPIQ. Low values on the y-axis (close to 0) indicate that majority of the vents are open. High values (close to 1) indicate that most of vents are closed.



(b) Same as **(a)**, but only for those samples where the 84$^{\text{th}}$ percentile of the predicted optimal MPIQ probability distribution function $\leq$ 50$^{\text{th}}$ percentile of the predicted reference MPIQ PDF. These are the same samples used to plot Figure 10d. We call these *robust* samples.



(c) Histogram of Hamming distances of predicted optimal vent configurations. This is the histogram of the y-axis values in **(a)**.



(d) Histogram of Hamming distances of predicted optimal vent configurations for the *robust* subset. This is the histogram of the y-axis values in **(b)**.

Figure 11. Distribution of predicted optimal vent configuration. Distribution is measured as Hamming distance from the all-open baseline.

a *robust* subset. On average, for all 10 train-test splits, this results in 62 *robust* samples. We emphasize that that this is a second, distinct sampling, with no guarantees of overlap with the 100 randomly selected samples used to plot Figure 9a. If we were to mis-actuate the vents and decrease the MPIQ, there would be a negative effect on the downstream science applications. To mitigate this risk here we chart only those instances where, if our models were to be put into production, we would be very confident in directing the telescope operator to move the vents according to our predictions. From Figure 9b we observe that all-open is not the optimal configuration in most situations. This is true even at lower measured MPIQ values. In fact, significant gains in MPIQ can be realized by switching each observation from the all-open configuration to the best configuration (from out limited choice of in-distribution and therefore "viable" configurations) is significant. When we consider higher MPIQs ($\sim\geq 1''$), the optimal configuration is likely to be all-close.

In Figure 10 we present results on the improvement in MPIQ predicted by our MDN given the (hypothetically) optimal vent configurations selected from the restricited set of in-distribution configurations selected by the RVAE. Figures 10a and 10b plot the improvement versus measured IQ, while Figures 10c and 10d plot the improvement versus predicted and calibrated median MPIQ. Further, the two right-hand plots, Figures 10b and 10d plots the improvement for *robust* sub-samples (the samples from the 84$^{\text{th}}$ quantiles) discussed in the last paragraph. We note two differences here, compared to Figure 9. First, we use all $\sim$ 600 viable test samples (for Figures 10a and 10c) in a given train-test split, and not sub-sample of 100; in Figure 9 we were forced to sub-sample due to space constraints. Second, here we show the difference in predicted optimal MPIQs with respect to both the predicted all-open MPIQs, and the measured MPIQs. For Figures 10a and 10c, we pick the optimal vent configuration for each of the $\sim$ 600 viable samples, use the MDN to derive calibrated predictions of MPIQ, and plot the difference between predicted calibrated optimal median and predicted calibrated median for the corresponding test sample with all twelve vents open. For Figures 10b and 10d, we do the same, but only for the *robust* samples from the $\sim$ 600 samples. We observe

that as we consider larger measured MPIQ values there is increased advantage to optimizing the vent configuration.

In the four sub-figures in Figure 11 we visualize the distribution of Hamming distances for the optimal vent configurations chosen for the samples in Figures 10. Either from the density of points plotted in Figure 11a, or from the distribution of Hamming distances from the all-open configuration plotted in Figure 11c, we observe that most optimal vents configurations are close to either all-open or all-closed. This should not come as a surprise. As discussed earlier, most vent configurations that are *not* close to either the all-open or all-closed configurations were OoD and so were filtered out of the "viable" set of configurations for which we consider the MPIQ predictions. At the risk of repeating some of the broader context provided earlier, the fact that the intermediate vent configurations are OoD is a result of the way the observatory has been operated to date. Most often the vents have been configured either all-open or all-closed, to a large degree because observers have had no reasoned methodology to follow to choose alternate configurations. The training data we have access to therefore clusters around the all-open and all-closed configurations. In a sense then, Figure 11 is another illustration of a main motivation for our work; we want to expand the range of options for the observers so they can better tune observatory performance.

An important observation from Figures 11a and 11c is that the model predicts about 60% of samples would have resulted in improved MPIQ had a different setting been chosen. Figures 11b and 11d, which are the predictions for the "robust" subset discussed earlier are even more definitive – a ful 85% of samples would have benefitted from a different vent configuration. About half of the adjustments would have be to close a *single* vent, while the other half would have closed all 12 vents. Only a smattering of predictions falls between these two choices. Of course, as just discussed, the intermediate range is mostly OoD.[15] That said, the peakiness at one-closed-vent and at all-12-closed-vents is, for us, a strong indication that the true optimal configuration lies somewhere in the middle. Not till we can collect additional data on this intermediate range to bring it in distribution will we be able to make robust predictions in that range that we can use to advise – with confidence – how the observer might more productively operate the telescope.

5.3 Quantifying feature importance in prediction of MPIQ

As a final contribution, we quantify the relative importance of each of the 119 features in predicted median MPIQ. By leveraging the integrated gradients technique (Janizek et al. 2020), we can attribute a *Shapley* score to each feature for each sample. This score measures the linear change in the predicted output (with respect to the average of the MPIQs across all samples in the training set, which is called the 'offset') that is induced by a small change in any given feature (i.e., the gradient). A positive score for a feature f in sample x implies that in x, f acts to increase the predicted MPIQ, while a negative score points to f's role in decreasing the predicted MPIQ. The larger the magnitude is of this Shapley" score, the bigger the role of f in determining MPIQ for x. We average such scores for all 119 features across all samples of interest, and plot *attribution plots* in Figure 12. In Figure 12a, we carry out this exercise for all test samples for each train-test split of $\mathcal{D}_{\mathcal{F}_S,\mathcal{S}_S}$, and collate the results from all 10 cycles (assuming the number of splits is 10). In Figure 12b, we repeat this exercise, but only for the significantly smaller test set samples ($\sim$ 600 per train-test split) that are in-distribution (ID).

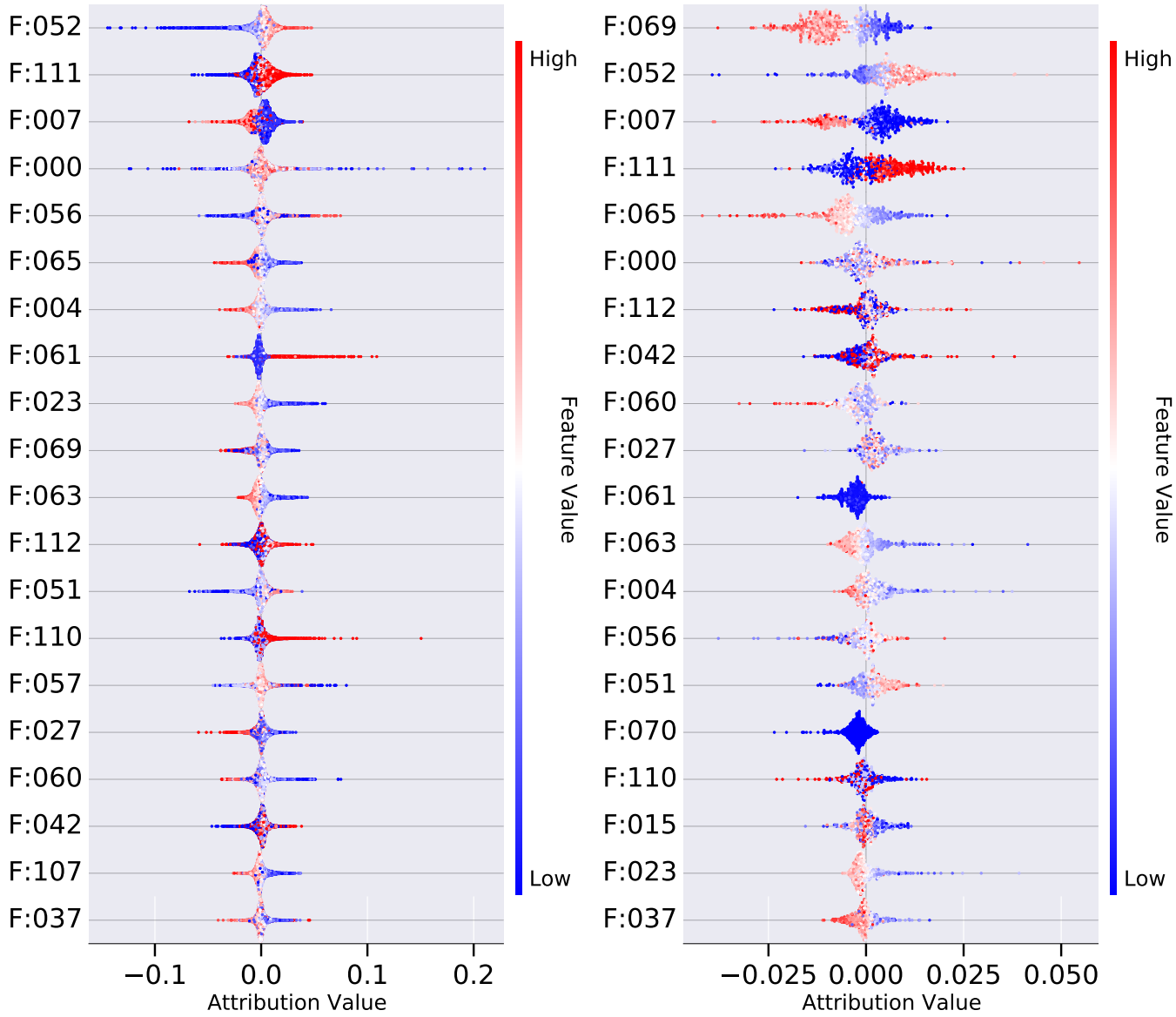
Concentrating on in-distribution samples (12b), we can see that the most important features in the prediction of MPIQ can be grouped roughly into three groups, related respectively to dome convection, seasonal variations and filter central wavelengths. In the first group, additional dome turbulence can be sourced by air convection, which is sourced by (positive) internal temperature gradients within the dome, with respect to altitude in the dome. Therefore, MPIQ is expected to increase with the mirror temperature that acts as a source of convection, and decrease with the temperature of the upper structures of the telescope (truss), that tends to reduce the temperature gradient. This is precisely what we see in the attribution plots of the two most important features. The third feature in importance shows a correction of the predicted MPIQ with respect to the $\lambda^{1/5}$ law used in Equation 3, with a predicted MPIQ larger at smaller wavelength compared to the theoretical scaling. Finally, the fourth ranked feature shows the seasonal variation of image quality, with better average seeing during the summer months due to more clement weather. Further features in the list can most of the time be attributed to one of the groups described above.

6 CONCLUSIONS AND FUTURE WORK

In this paper we present what we envisage to be the first in a series of studies that will ultimately lead to dynamically optimized scheduling at the Canada-France-Hawaii Telescope. We have initiated that program herein by developing machine-learning based data-driven methods of image quality (IQ) prediction. We present results for two models. The first is a feed-forward mixture density network (MDN) used in conjunction with a robust variational autoencoder. We trained both on a new dataset that comprises eight years of data collected at CFHT since the installation of the dome vents. The MDN produces probabilistic predictions of image quality, while the autoencoder estimates the marginal distribution of the data. On average, IQ can be predicted to within 0.07" accuracy based on environmental conditions and telescope operating parameters. By varying the configuration of the dome vents (in an in-distribution way) in response to environmental conditions, our model predicts that IQ can be improved by about 10% over historical patterns, with the gains increasing when the nominal IQ value is large. For SNR-based observations, this represents gains of up to 10-15%. These gains, in turn, can be equated to approximately 1M USD in operating costs per year of SNR-based observing. Such gains would be realized in the form of additional observations made and experiments conducted; additional science accomplished.

We see several important avenues for further inquiry. Perhaps most immediate, the improvements in IQ that we present are predicted by extrapolating over hypothetical vent configurations. While the uncertainties predicted by our model suggests that these out-of-distribution predictions are robust, we need to verify our predictions by collecting additional observation data in these operating regimes. By collecting such data we will be able to extend our model and robustly predict IQ for the more intermediate vent configurations

[15] At this point is is very helpful to refer back to Figure 8b and observe that 'single-vent-closed' is the third most frequent vent setting in the training set. This explains why our model finds that this type of close-a-single-vent adjustment is in-distribution. Further, the $\sim$ 38% prevalence of one-closed-vent configurations in the robustified results of Figures 11b and 11d tells us that this option is of great use in improving the (robustly predicted) MPIQ resulst plotted in Figures 10b and 10d.



(a) Attribution plot for all samples. This gives a bird's eye view of which features the model sees as most predictive.

(b) Attribution plot for in-distribution (ID) samples with all vents open. This enables us to isolate those features which make samples ID.

Figure 12. Attribution (a.k.a. summary) plots. By using expected gradients (Janizek et al. 2020), we obtain the impact of each feature on the mixture density network's predicted output. These are then collated for all samples in the test set (for a given fold), and collated again for all test sets from all folds. **(a):** Using *all* ~ 60000 samples. **(b):** Using only the ~ 6000 in-distribution (ID) samples for which there exist vent configurations that result in lower predicted median MPIQ than their counterparts with all 12 vents open.

Table 6. Most predictive features identified in Figure 12.

Abbreviation	Feature	Abbreviation	Feature
F:000	Barometric pressure	F:057	Rear observing room air temperature
F:004	Catwalk temperature, north	F:060	Thrust bearing surface temperature, south beam
F:007	Filter central wavelength	F:061	Top ring air temperature, east
F:015	Current altitude	F:063	Top ring air temperature, west
F:023	Dome top temperature	F:065	Truss surface temperature, north halfway-up
F:027	Dome wall temperature, west	F:069	Truss surface temperature, west halfway-up
F:037	Fourth floor crawlspace air temperature	F:070	Vent L1
F:042	sin(hour of day)	F:107	Weather tower temperature
F:051	Mirror surface temperature, east underside	F:110	Weather tower wind speed
F:052	Mirror surface temperature, south underside spigot	F:111	cos(week of year)
F:056	Observing room air temperature	F:112	sin(week of year)

"half-way" between all-open and all-closed. In doing so we aim, finally, to realize the full utility of the dome vents.

Second, in this study we have treated each data record as an independent sample. In reality of course, exposures are temporally related. Numerous exposures are collected each night. By treating data records as independent, we do not leverage what we anticipate are quite important temporal relationships extant in the data. By incorporating temporal models, our aim is be to be able to produce real-time robust forecasts of IQ some five-to-twenty minutes into the future. The realization of such capabilities will enable the adaptive reorganization of a nominal observation schedule, the real-time scheduling protocols we mention in the introduction.

Connected to the second point, in this paper we work exclusively with data records from MegaCam. Going forward we plan to augment our data set with records from other CFHT instruments. While MegaCam's IQ measurements are the most accurate, other instruments also measure IQ with acceptable accuracy, and equally importantly, operate when MegaCam is offline. CFHT schedules instruments in blocks or "runs" of several consecutive nights, e.g., swapping out instruments twice a month according to their sensitivity to moonlight. The consequence of this is that training data from MegaCam is not temporally contiguous. This makes it more difficult to use in training a scheduler.

Fourth, we do not currently take into account the physical locations of the different sensors. While discarding this spatial information made modeling and data analysis easier for this initial study, it leaves out useful side information that can connect the placement of sensors and their relative values. Going forward we will incorporate such information into our models.

Fifth, the approach we take to *post-hoc* calibration of epistemic uncertainties – CRUDE (Zelikman & Healy 2020) – has its own set of limitations. While state-of-the-art in terms of improving sharpness and calibration, CRUDE implicitly assumes a symmetric distribution of uncertainty. In our context this means we do not fully leverage the *asymmetric* uncertainties output by our β posteriors. CRUDE also assumes that, once normalized (by their standard deviations), all errors are drawn from the same distribution. Therefore CRUDE weights all data points equally. In our context one implication is that while calibrating the probability distribution function (PDF) for a test sample with nominal (true) IQ of $2''$, which is near the right tail of the distribution (see the red curve in the left sub-figure of Figure 2), we are strongly impacted by samples with nominal IQ values near 0.6 arc-seconds, which is the mode of the distribution. This uniformity of treatment is not ideal. We plan to address this shortcoming in the future.

Sixth, we note that we are cautious about the thresholding method we apply to detect out-of-distribution samples (see Figure 8a). In this work we take the 95^{th} percentile of the pseudo marginal log-likelihoods for the training set samples as the out-of-distribution threshold. However, this is an *ad-hoc* choice based on intuition. We do not claim that it is the optimal method to filter out out-of-distribution samples. We also show that log-likelihood regret is a more accurate metric than is log-likelihood when aiming to separate the two types of distributions. We refrain from using regret in this work due to practical concerns about run-time. In future work, we will leverage distributed computing to integrate this superior metric into our pipeline, and we will explore more principled ways to set the threshold.

Finally, in a slightly different direction, we note that a subset of the authors are collaborating with a concurrent and complimentary study of dome seeing at CFHT. In that study direct measurements of local in-dome optical turbulence are being collected using AIRFLOW instruments Lai et al. (2019). AIRFLOW sensors are always-on optical turbulence sensors and, as discussed in the introduction, turbulence is highly correlated with instrument IQ, the metric of interest herein. The current work informs the AIRFLOW study in that it can provide insight into which locations sensors should be placed. Conversely, data from the AIRFLOW study can provide a new data stream for the current study. Taken together, these two studies will offer unique insights into the nature of dome seeing and ways that effects which degrade seeing can be mitigated.

ACKNOWLEDGEMENTS

This research was based on observations obtained at CFHT, which is operated from the summit of Maunakea by the National Research Council of Canada, the Institut National des Sciences de l'Univers of the Centre National de la Recherche Scientifique of France, and the University of Hawaii. The authors wish to recognize and acknowledge the very significant cultural role that the summit of Maunakea has always had within the indigenous Hawaiian community. S.D. and S.G. thank CFHT for hosting them at the beginning of the project. S.G. thanks Eric Zelikman (Stanford University), Haleh Akrami (University of Southern California), Joseph Jazinek (Department of Computer Science, University of Washington), and Dr. Yusuke Nomura (Department of Radiation Oncology, Stanford University) for productive e-mail correspondence. S.P. thanks Olivier Lai for fruitful discussions. The authors would like to thank the CFHT staff, and Stephen Gwyn at the Canadian Astronomy Data Centre. Y.S.T. is grateful to be supported by the NASA Hubble Fellowship grant HST-HF2-51425.001 awarded by the Space Telescope Science Institute.

REFERENCES

Akrami H., Joshi A. A., Li J., Aydore S., Leahy R. M., 2019, arXiv preprint arXiv:1905.09961
Akrami H., Aydore S., Leahy R. M., Joshi A. A., 2020, arXiv preprint arXiv:2006.08204
Baril M., Benedict T., Thanjavur K., Salmon D., Vogiatzis K., Racine R., Breidenthal R., 2012, in Modeling, Systems Engineering, and Project Management for Astronomy V. p. 844903
Beckers J. M., 1993, Annual review of astronomy and astrophysics, 31, 13
Bengio Y., Delalleau O., Simard C., 2010, Computational Intelligence, 26, 449
Bishop C. M., 1995, Neural Networks for Pattern Recognition. Oxford University Press, Inc., USA
Boulade O., et al., 2003, in Iye M., Moorwood A. F. M., eds, Vol. 4841, Instrument Design and Performance for Optical/Infrared Ground-based Telescopes. SPIE, pp 72 – 81, doi:10.1117/12.459890, https://doi.org/10.1117/12.459890
Bracher J., Ray E. L., Gneiting T., Reich N. G., 2021, PLOS Computational Biology, 17, e1008618
Breiman L., 2001, Machine learning, 45, 5
Choi S., Lee K., Lim S., Oh S., 2018, in 2018 IEEE International Conference on Robotics and Automation (ICRA). pp 6915–6922
Collier M., Nazabal A., Williams C. K., 2020, arXiv preprint arXiv:2006.05301
Crabtree D., 2019, Canada's astronomy performance based on bibliometrics, doi:10.5281/zenodo.3756124, https://doi.org/10.5281/zenodo.3756124
Eguchi S., Kato S., 2010, Entropy, 12, 262
Friedman J. H., 2001, Annals of statistics, pp 1189–1232
Fu H., Li C., Liu X., Gao J., Celikyilmaz A., Carin L., 2019, arXiv preprint arXiv:1903.10145

Futami F., Sato I., Sugiyama M., 2018, in International Conference on Artificial Intelligence and Statistics. pp 813–822
Gal Y., 2016, in Uncertainty in deep learning. University of Cambridge
Ghosh A., Basu A., 2016, Annals of the Institute of Statistical Mathematics, 68, 413
Gilda S., et al., 2020, arXiv e-prints, p. arXiv:2011.03132
Gilda S., Lower S., Narayanan D., 2021a, arXiv e-prints, p. arXiv:2101.04687
Gilda S., Lower S., Narayanan D., 2021b, mirkwood: SED modeling using machine learning (ascl:2102.017)
Gneiting T., Raftery A. E., 2007, Journal of the American statistical Association, 102, 359
Haldar M., et al., 2019, in Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. pp 1927–1935
Hinton G. E., Salakhutdinov R. R., 2006, science, 313, 504
Hoffman M. D., Johnson M. J., 2016, Workshop in Advances in Approximate Bayesian Inference, NeurIPS
Ioffe S., Szegedy C., 2015, arXiv preprint arXiv:1502.03167
Izmailov P., Podoprikhin D., Garipov T., Vetrov D., Wilson A. G., 2018, arXiv preprint arXiv:1803.05407
Janizek J. D., Sturmfels P., Lee S.-I., 2020, arXiv preprint arXiv:2002.04138
Jordan M. I., Ghahramani Z., Jaakkola T. S., Saul L. K., 1999, Machine Learning, 37, 183
Kingma D. P., Ba J., 2015, in Bengio Y., LeCun Y., eds, 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings. http://arxiv.org/abs/1412.6980
Kingma D. P., Welling M., 2013a, arXiv preprint arXiv:1312.6114
Kingma D. P., Welling M., 2013b, arXiv preprint arXiv:1312.6114
Kingma D. P., Welling M., 2019, arXiv preprint arXiv:1906.02691
Kuleshov V., Fenner N., Ermon S., 2018, arXiv preprint arXiv:1807.00263
Lai O., Withington J. K., Laugier R., Chun M., 2019, Monthly Notices of the Royal Astronomical Society, 484, 5568
Lakshminarayanan B., Pritzel A., Blundell C., 2017, in Advances in neural information processing systems. pp 6402–6413
Li B., Wu F., Weinberger K. Q., Belongie S., 2019, arXiv preprint arXiv:1907.04312
Li B., Wu F., Lim S.-N., Belongie S., Weinberger K. Q., 2020, arXiv preprint arXiv:2002.11102
Lyman R., Cherubini T., Businger S., 2020, 2020MNRAS.496.4734L
Malinin A., Prokhorenkova L., Ustimenko A., 2021, ICLR 2021
Milli J., et al., 2019, arXiv preprint arXiv:1910.13767
Misra D., 2019, arXiv preprint arXiv:1908.08681
Nixon J., Dusenberry M., Zhang L., Jerfel G., Tran D., 2019, arXiv, pp arXiv–1904
Quinlan J. R., 1986, Machine learning, 1, 81
Racine R., 1984, vlti.conf..235R
Racine R., Salmon D., Cowley D., Sovka J., 1991, Publications of the Astronomical Society of the Pacific, 103, 1020
Racine R., Salman D., Bauman S., Mahoney B., Cuillandre J.-C., 2018, CFHT Internal Report
Reddi S., Zaheer M., Sachan D., Kale S., Kumar S., 2018, in Proceeding of 32nd Conference on Neural Information Processing Systems (NIPS 2018).
Rezende D. J., Mohamed S., Wierstra D., 2014, in International conference on machine learning. pp 1278–1286
Roddier F., 1988, Applied Optics, 27, 1223
Salmon D., Cuillandre J.-C., Barrick G., Thomas J., Ho K., Matsushige G., Benedict T., Racine R., 2009, Publications of the Astronomical Society of the Pacific, 121, 905
Sarazin M., Roddier F., 1990, A&A, 227, 294
Skidmore W., Els S., Travouillon T., Riddle R., Schöck M., Bustos E., Seguel J., Walker D., 2009, Publications of the Astronomical Society of the Pacific, 121, 1151
Smith L. N., 2017, in 2017 IEEE Winter Conference on Applications of Computer Vision (WACV). pp 464–472, doi:10.1109/WACV.2017.58
Tokovinin A., Vernin J., Ziad A., Chun M., 2005, Publications of the Astronomical Society of the Pacific, 117, 395
V. I. Tartarskiĭ R. A. Silverman (ed.) 1961, Wave Propagation in a Turbulent Medium. University of Michigan: McGraw-Hill Books, Belmont, MA
Wu Y., He K., 2018, in Proceedings of the European conference on computer vision (ECCV). pp 3–19
Xiao Z., Yan Q., Amit Y., 2020, arXiv preprint arXiv:2003.02977
Zelikman E., Healy C., 2020, arXiv preprint arXiv:2005.12496

APPENDIX A: FIGURES

In this appendix we present a few figures that we anticipate will help the reader better understand our implementation of the machine learning techniques we use. In particular, in Figure A1 we illustrate our MDN training process. In Figure A2 we illustrate the approach taken to identify in-distribution vent configurations amongst all possible 2^{12} configurations. Our predictions are restricted to only in-distribution vent configutations. Finally, in Figures A3 and A4 we present some details on how we select our learning rate.

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

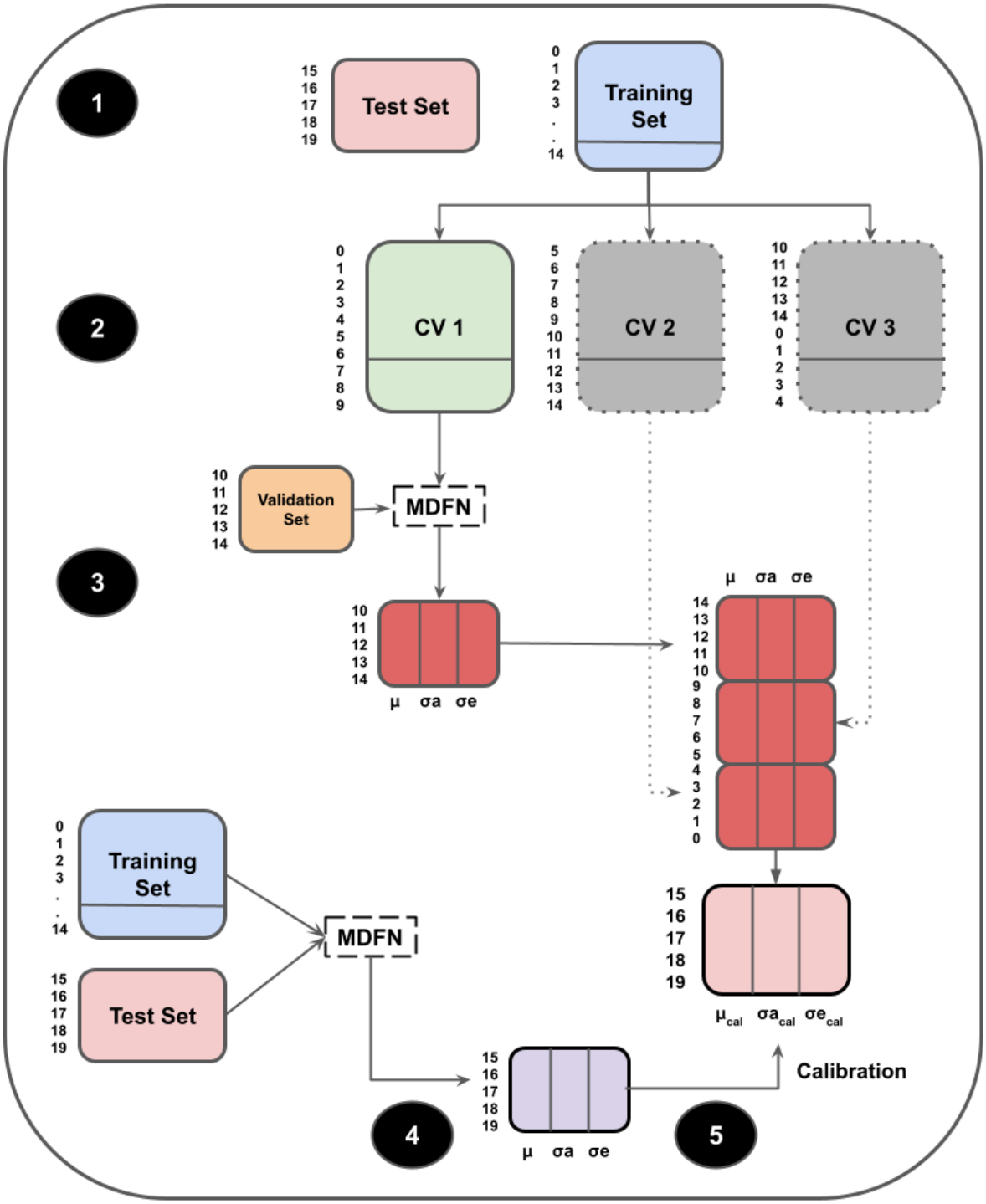


Figure A1. To save space, we only illustrate **Step 1:** We divide the input data set $\mathbf{X}$ into $N_{\mathrm{CV},1} = 10$ cross-validation folds, at any point referring to the collation of $N_{\mathrm{CV},1} - 1$ of them as $\mathbf{X_{TRAIN}}$, and the remaining fold as $\mathbf{X_{TEST}}$. We repeat this $N_{\mathrm{CV},1}$ times to cover all samples in $\mathbf{X}$, but depict only one such iteration here for illustration. **Step 2:** We sub-divide $\mathbf{X_{TRAIN}}$ into $N_{\mathrm{CV},2} = 3$ CV folds. Same as before, $N_{\mathrm{CV},2} - 1$ folds are collated while the remaining fold, referred to as the validation set $\mathbf{V}$, is set aside. **Step 3:** Each of the $N_{\mathrm{CV},2}$ CV folds is used to create $N_{\mathrm{bags}} = 3$ 'bags' by randomly shuffling its data and picking the same number of samples with replacement. The Mixture Density Feedforward Network (MDFN), with architecture visualized in Figure 4, is trained on one of such folds, and predictions on the validation set give us the 16^{th}, 50^{th}, and 68^{th} predictions, plus the epistemic uncertainty per sample in $\mathbf{V}$. This process is repeated $N_{\mathrm{CV},2} - 1$ more times to get predictions for all samples in $\mathbf{X_{TRAIN}}$. **Step 4:** The MDFN is now trained on the entire training set, and predictions collected for samples in $\mathbf{X_{TEST}}$. **Step 5:** Finally, we use CRUDE (Zelikman & Healy 2020), and the predictions from Step 3 to calibrate the predicted values from Step 4.

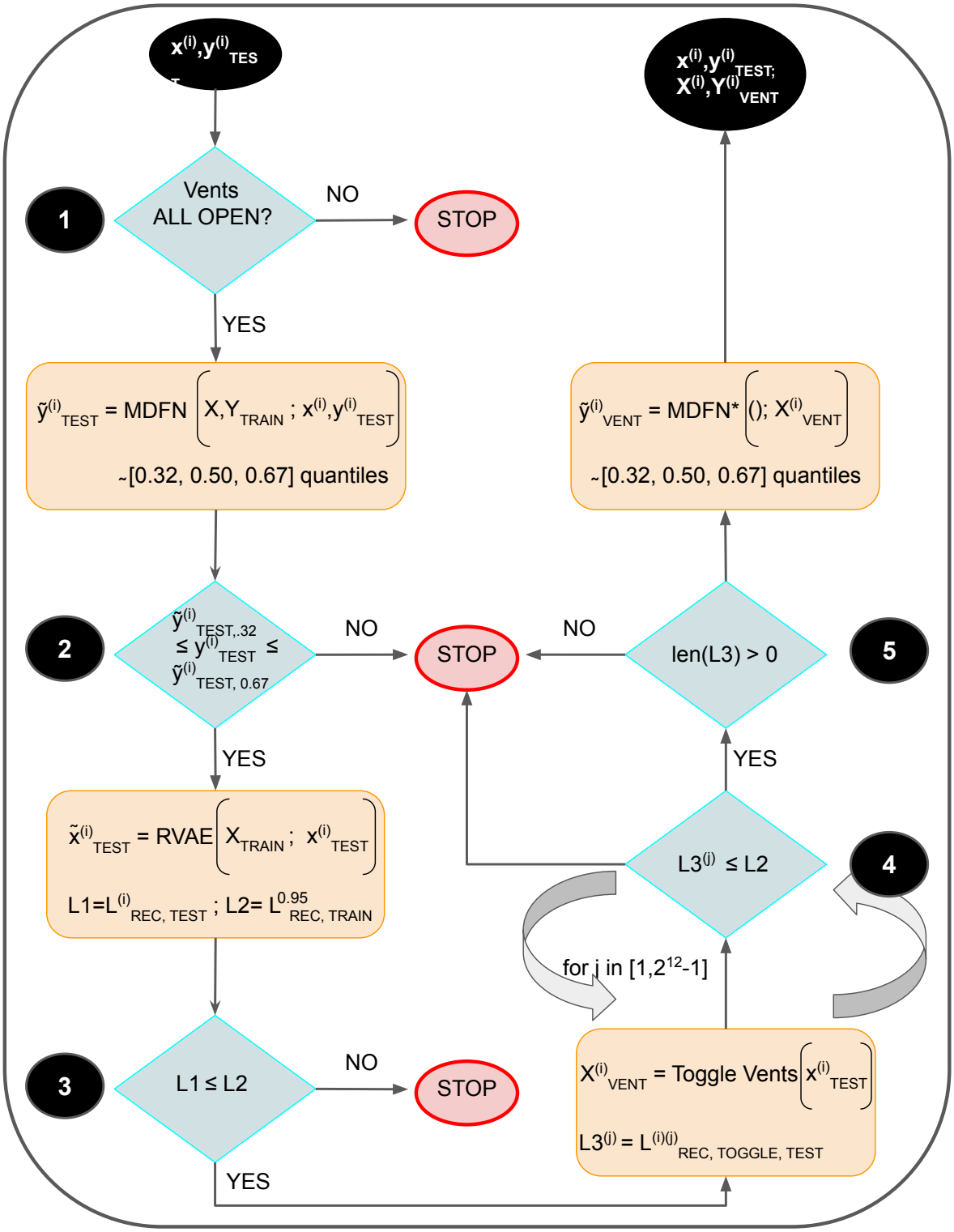
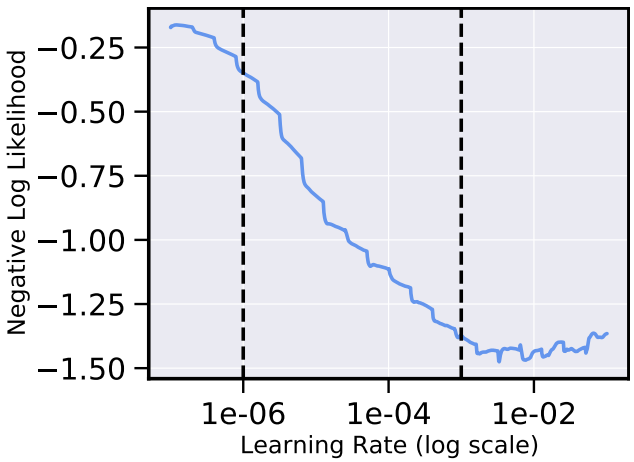
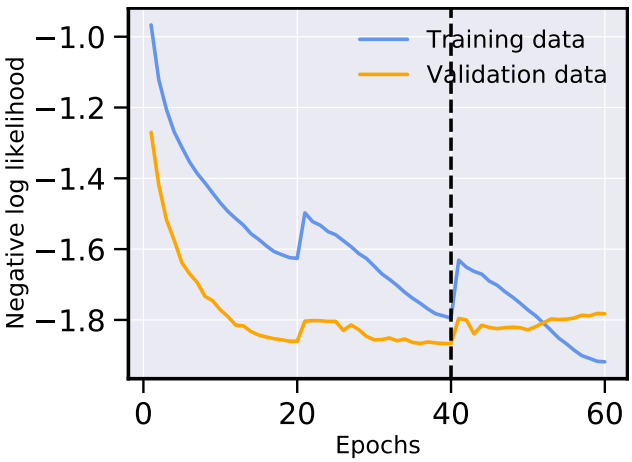


Figure A2. Workflow for identifying acceptable, in-distribution (ID) vent configurations, for each sample in the test set. **First**, we isolate samples with all twelve vents open. **Second**, we further pick only those samples for which the predicted 68^{th} and 16^{th} quantiles, generated by adding the epistemic and aleatoric uncertainties in quadrature, envelope the true IQ. 'MDFN' stands for Mixture Density Feedforward Network, see Figure A1. **Third**, we use the training set with our robust VAE, calculate the 95^{th} percentile of the reconstruction loss, and use that as an upper cut-off to separate in-distribution test samples from out-of-distribution (OoD) ones. **Fourth**, for the test samples thus filtered, we generate $2^{12} - 1$ samples by toggling the 12 vents into open and close positions, but skip the all-open configuration since that is the base case. From these hypothetical samples, we find the ID ones by repeating the procedure of step 3. **Finally**, we throw out those test samples for which none of the hypothetical cases passed the cut-off test.

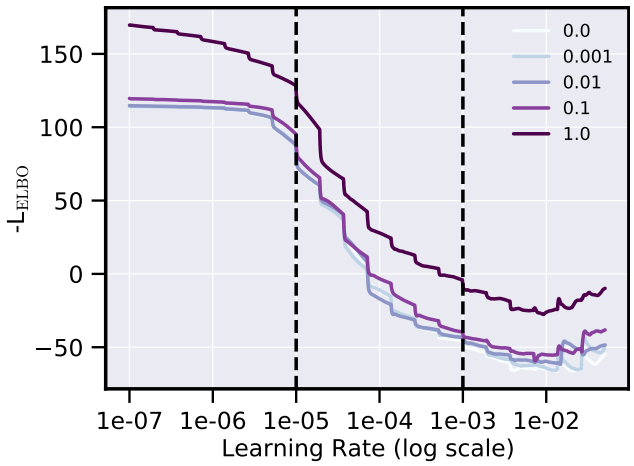


(a) Choosing lower and upper learning rates for the cyclic learning rate scheduler, as described in Section 4.8. The vertical lines indicate the chosen limits of 10^{-6} and 10^{-3}.

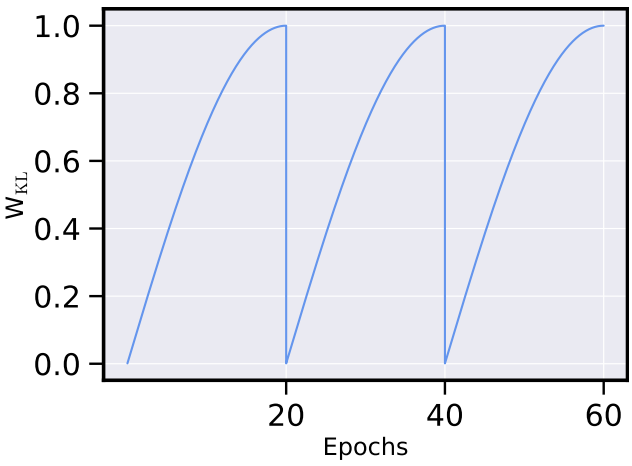


(b) Training and validation curves for one of ten folds when using the Mixture Density Network, as described in Section 4.8. The vertical line indicates the epoch of minimum validation loss.

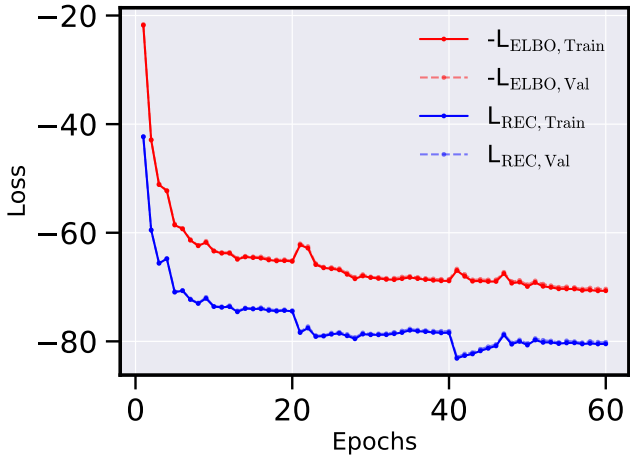
Figure A3. Loss as a function of learning rate and epochs for one of ten folds when training using the MDN. In **(b)**, the training loss is higher than the validation loss owing to the MoEx augmentation (Li et al. 2020), as we explain in Section 4.1.



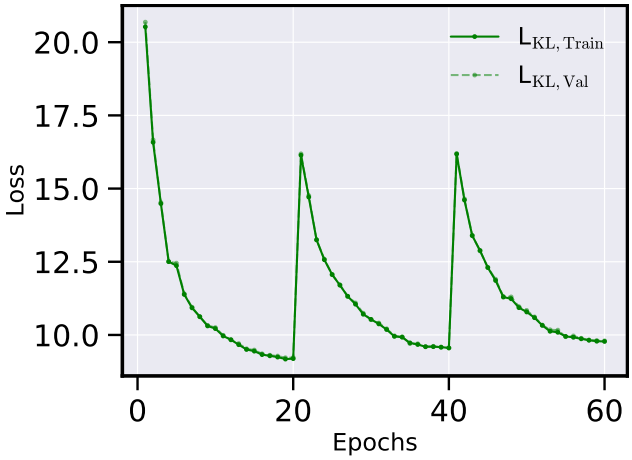
(a) Choosing lower and upper learning rates for the cyclic learning rate scheduler. Since we also anneal W_{KL}, it is important to find LR limits that encourage quick convergence over the entire domain of W_{KL}.



(b) Annealing W_{KL} from 0 to 1 in each cycle of 20 epochs.Finding optimal β for the robust variational autoencoder. This prevents L_{KL} from collapsing to 0, and is based upon findings of Fu et al. (2019).



(c) Training and validation curves for the total loss and the reconstruction loss for one of ten folds.



(d) Training and validation curves for the KL-divergence loss for one of ten folds.

Figure A4. Losses as a function of learning rate, W_{KL}, and epochs, for one of ten folds when training using the RVAE. While the MDN has a single loss (see Figure A3), the RVAE has two individual losses. In conjunction these form the final loss that is minimized by mini-batch gradient descent ($-L_{ELBO} = L_{REC} + L_{KL}$, see Section 4.3). Different from Figure A3b, **(c)** and **(d)** here are not 'live' plots, but constructed once the RVAE has been fully trained.