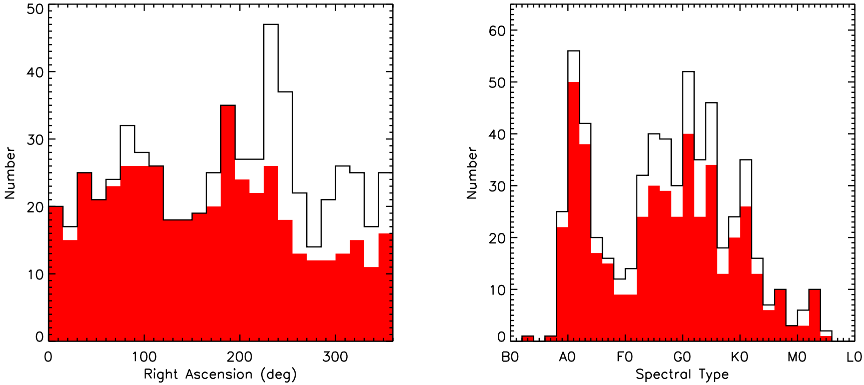
Mining the GPIES Database

[SPIE: 10703-17]

Dmitry Savransky
Jacob Shapiro, Vanessa Bailey, Robert De Rosa, Jason Wang, Jean-Baptiste Ruffio, Eric Nielsen, Melisa Tallis, Marshall Perrin
and the GPIES Team

June 10, 2018

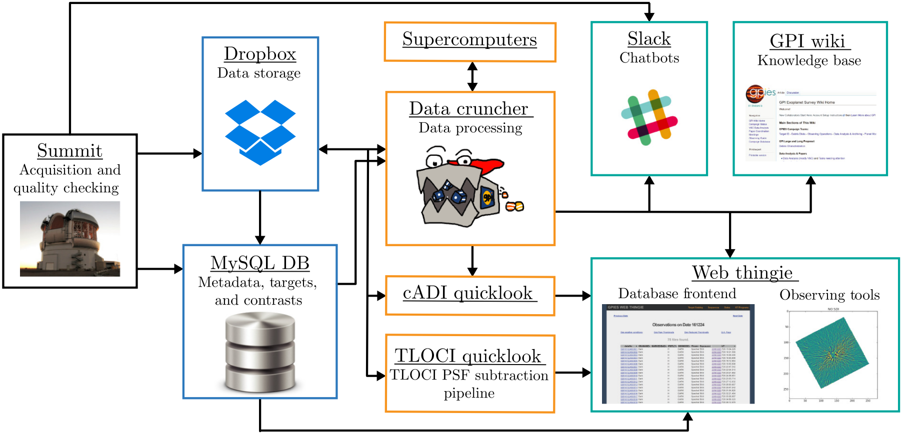
The Gemini Planet Imager Exoplanet Survey



484 targets observed to date

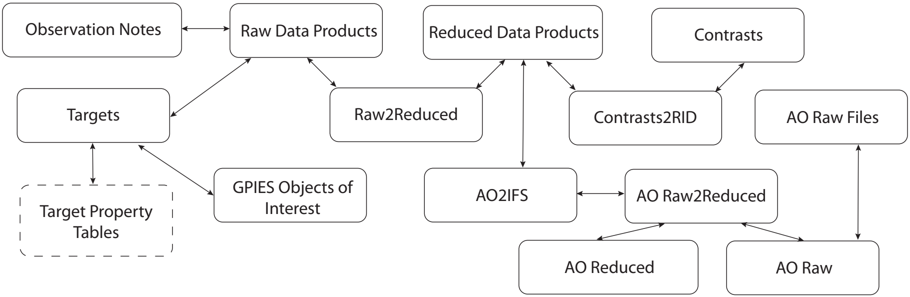
Figures courtesy of R. De Rosa

GPIES Campaign Data System



From: [Wang et al., 2018]

GPIES Database Schema



GPIES Database Contents

- 14 GB (metadata and ancillary products only)
- 136,151 IFS Raw Data Files (30,142 GPIES)
- 263,973 IFS Reduced Data Products
- 86,092 AO Raw Telemetry Files
- 86,325,374 Contrast Values

GPIES Database Contents

- 14 GB (metadata and ancillary products only)
- 136,151 IFS Raw Data Files (30,142 GPIES)
- 263,973 IFS Reduced Data Products
- 86,092 AO Raw Telemetry Files
- 86,325,374 Contrast Values

What can we do with all this data?

Previous Work

- Performance characterization of GPI's AO with IFS data [Poyneer et al., 2016, Bailey et al., 2016]
- GPI performance variation characterization with operating conditions [Tallis et al., 2018]
- See also: Tallis et al., this conference [10703-267]

Previous Work

- Performance characterization of GPI's AO with IFS data [Poyneer et al., 2016, Bailey et al., 2016]
- GPI performance variation characterization with operating conditions [Tallis et al., 2018]
- See also: Tallis et al., this conference [10703-267]

That's not what this talk is about

Here, we are only looking for purely data-driven results, with no specific physical modeling of underlying processes

Finding Correlations

For two random variables $\bar{x}, \bar{y}$:

- Pearson product-moment:

$$r_{\bar{x},\bar{y}} = \frac{E[(\bar{x} - \mu(\bar{x}))(\bar{y} - \mu(\bar{y}))]}{\sigma(\bar{x})\sigma(\bar{y})}$$

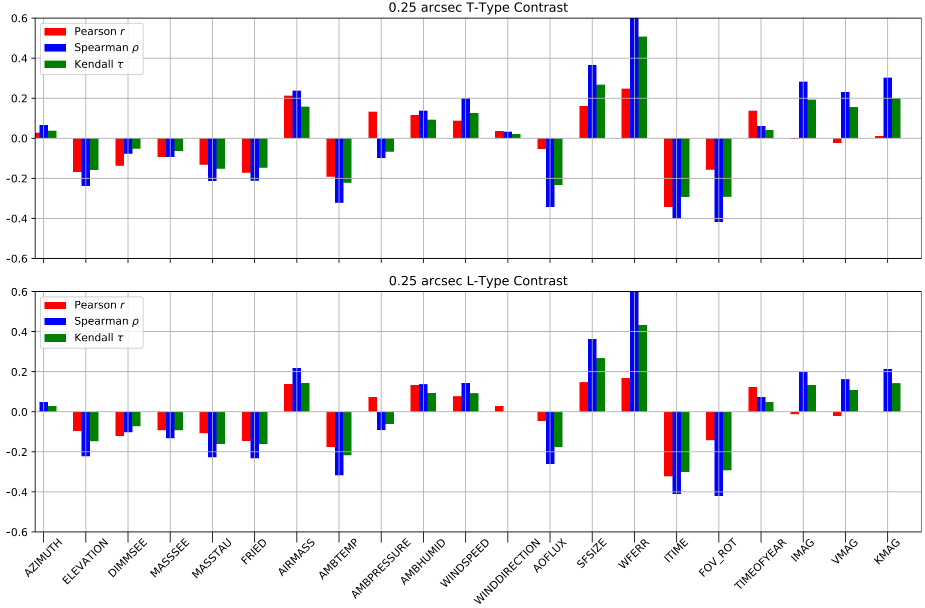
- Spearman rank correlation:

$$\rho_{\bar{x},\bar{y}} = r_{\mathrm{rank}\,\bar{x},\mathrm{rank}\,\bar{y}}$$

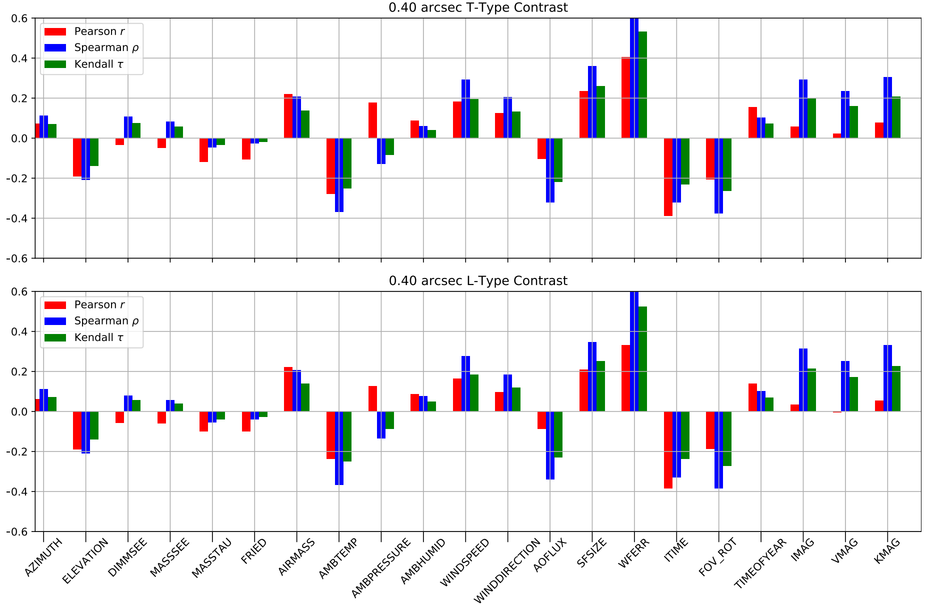
- Kendall rank correlation:

$$\tau = \frac{2}{n(n-1)} \left(\sum_{i \neq j} [((x_i > x_j) \& (y_i > y_j)) | ((x_i < x_j) \& (y_i < y_j))] \right.$$
$$\left. - \sum_{i \neq j} [((x_i < x_j) \& (y_i > y_j)) | ((x_i > x_j) \& (y_i < y_j))] \right)$$

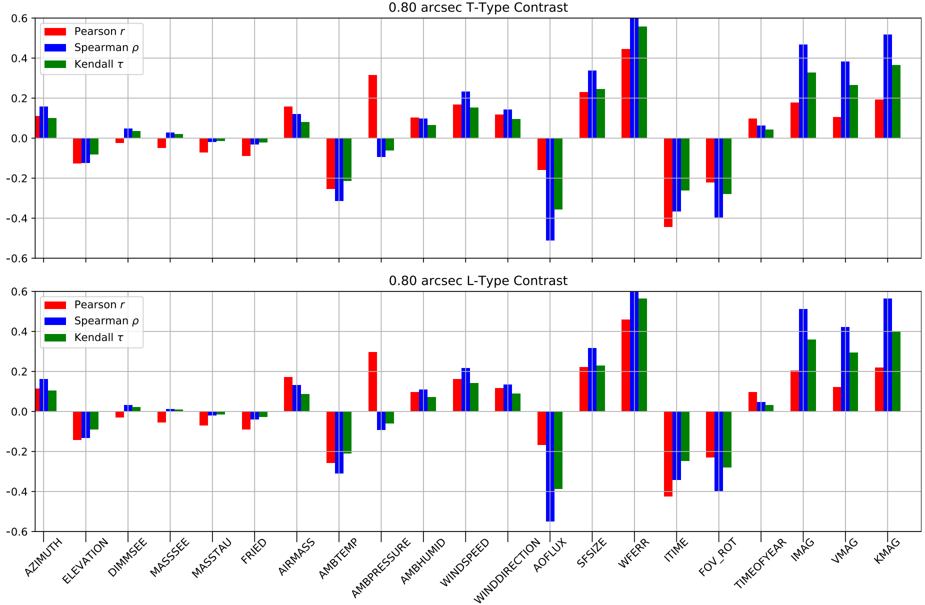
Contrast Correlations



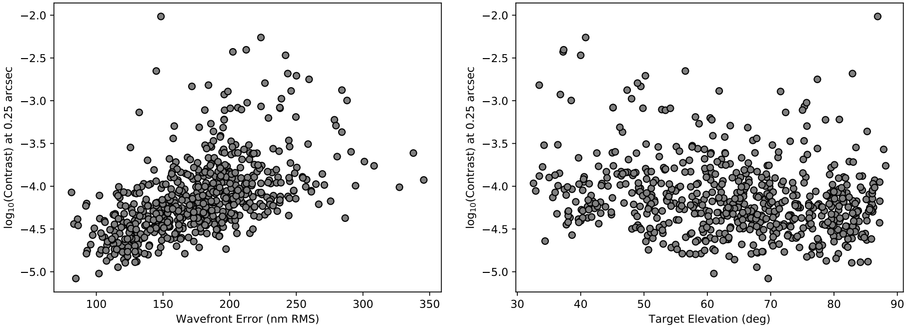
Contrast Correlations



Contrast Correlations



The Data is Noisy



Mixture Models

- A point y_i may belong to the "true" data or be considered an outlier drawn from a normal distribution $\sim(\mu_o, \sigma_o)$, governed by binary flag o_i:

$$p(y_i|\mathbf{x}_i, \sigma_i, \boldsymbol{\theta}, o_i, \mu_o, \sigma_o) = \frac{1}{\sqrt{2\pi\left(\sigma_i^2 + o_i\sigma_o^2\right)}} \exp\left(-\frac{\left[y_i - (1-o_i)f_{\boldsymbol{\theta}}(\mathbf{x}_i) - o_i\mu_o\right]^2}{2\left(\sigma_i^2 + o_i\sigma_o^2\right)}\right)$$

- The marginalized likelihood is then:

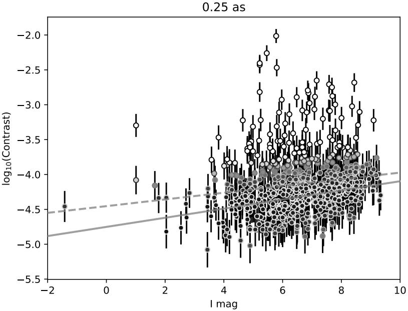
$$p(\{y_i\}_{i=1}^n | \{\mathbf{x}_i\}_{i=1}^n, \{\sigma_i\}_{i=1}^n, \boldsymbol{\theta}, \mu_o, \sigma_o) = \prod_{i=1}^n \left[O p(y_i|\mathbf{x}_i, \sigma_i, \boldsymbol{\theta}, o_i = 0) + (1-O) p(y_i|\mathbf{x}_i, \sigma_i, \boldsymbol{\theta}, o_i = 1)\right]$$

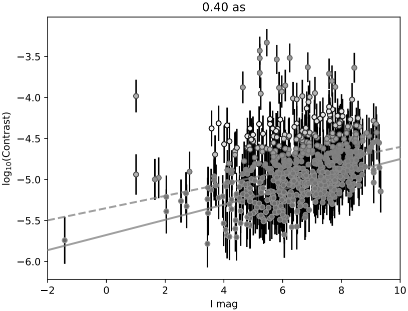
for

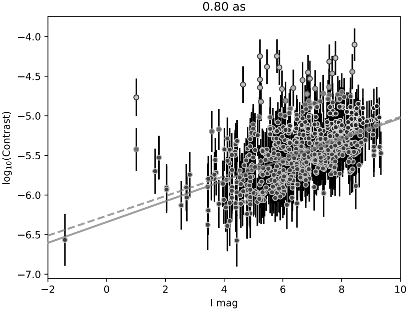
$$p(o_i) = \begin{cases} O & o_i = 0 \\ 1 - O & o_i = 1 \end{cases}$$

See: [Hogg et al., 2010, Hogg and Foreman-Mackey, 2017]

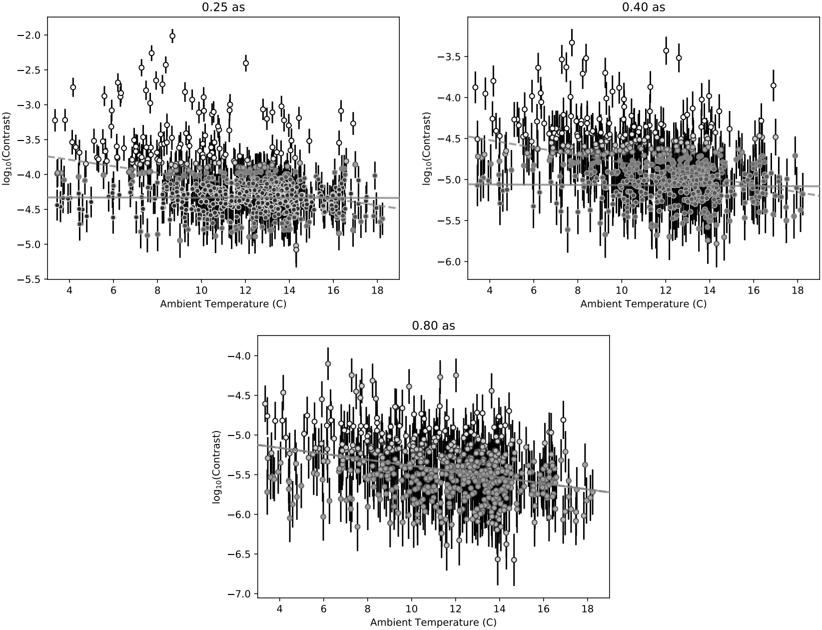
Linear Modelling (I-Magnitude)



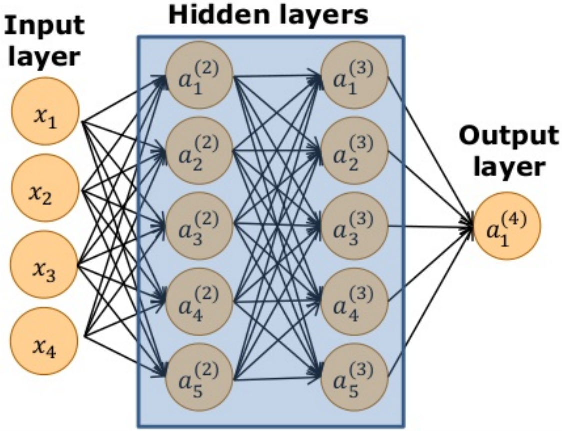




Linear Modelling (Ambient Temperature)



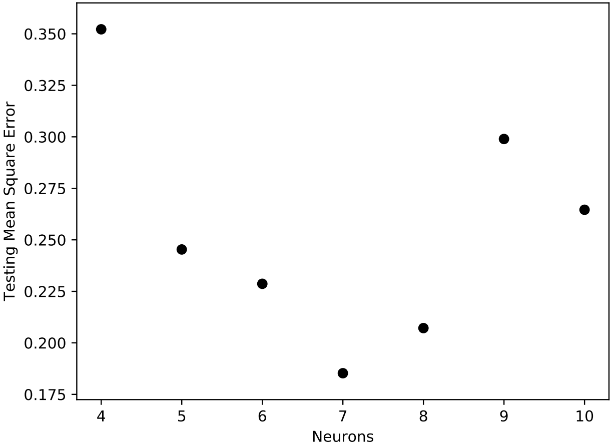
DNN Regression



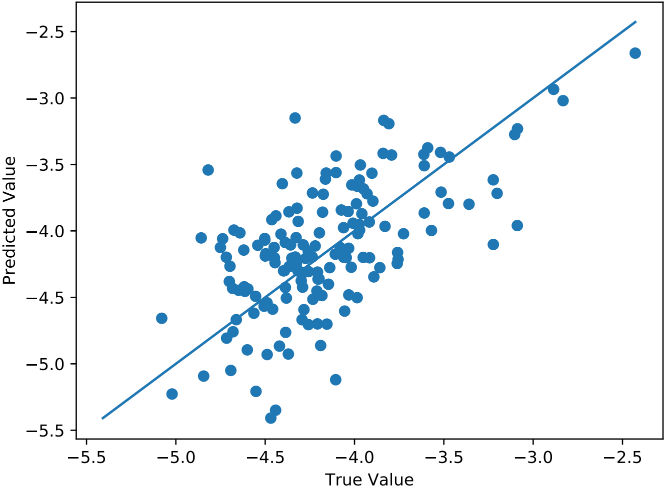
From: [G. Lion, 2016]

This work done entirely in TensorFlow r1.8.

Choice of Network

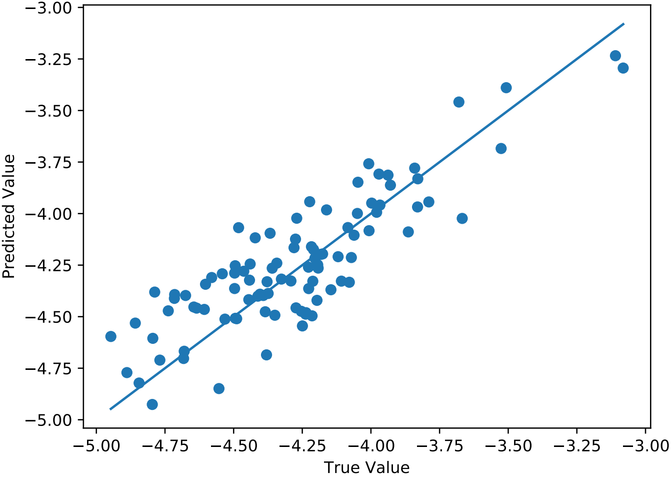


Single Layer, 8 Neuron, 9 Input Regression Network



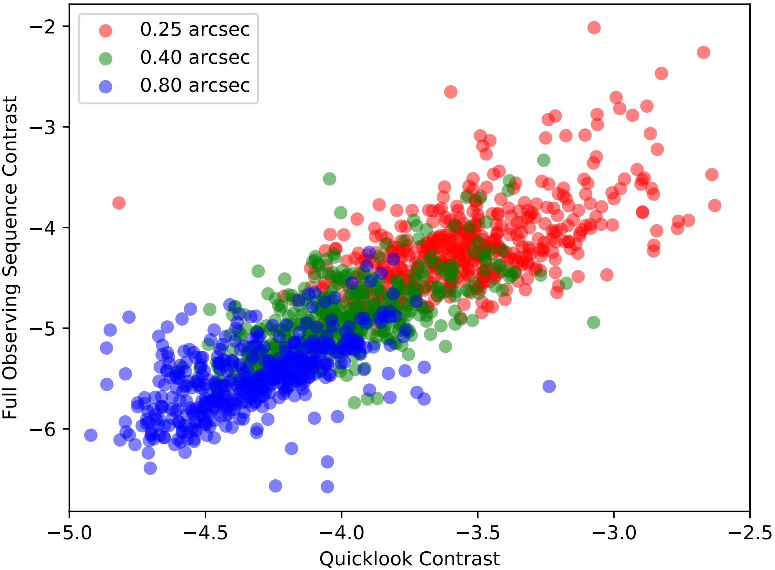
RMSE: 0.40

Two Layers, 16 Neuron, 6 Input Regression Network

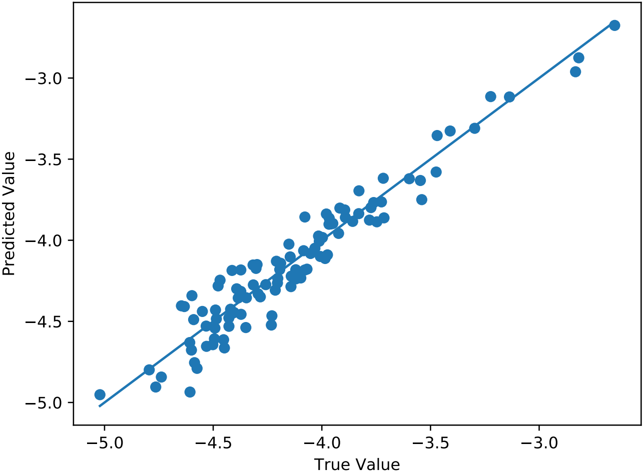


RMSE: 0.18

What Can We Say After the First Observation?



Three Layers, 60 Neuron, 22 Input Regression Network



RMSE: 0.11

Conclusions

- Jointly exploiting operational and science data metrics can lead to new discoveries, but is difficult if you don't have the proper infrastructure in place
- Polynomial models are likely insufficient to accurately describe performance variations given the large numbers of endogenous and exogenous factors in play
- Machine Learning is great, but it's hard to tell if you really have the right answer

References I

Bailey, V. P., Poyneer, L. A., Macintosh, B. A., Savransky, D., Wang, J. J., De Rosa, R. J., Follette, K. B., Ammons, S. M., Hayward, T., Ingraham, P., Maire, J., Palmer, D. W., Perrin, M. D., Rajan, A., Rantakyrö, F. T., Thomas, S., and Véran, J.-P. (2016).
Status and performance of the gemini planet imager adaptive optics system.
In *Proc. SPIE*, volume 9909, pages 99090V–99090V–15.

Hogg, D., Bovy, J., and Lang, D. (2010).
Data analysis recipes: Fitting a model to data.
Arxiv preprint arXiv:1008.4686.

Hogg, D. W. and Foreman-Mackey, D. (2017).
Data analysis recipes: Using markov chain monte carlo.
arXiv preprint arXiv:1710.06068.

Poyneer, L. A., Palmer, D. W., Macintosh, B., Savransky, D., Sadakuni, N., Thomas, S., Véran, J.-P., Follette, K. B., Greenbaum, A. Z., Ammons, S. M., Bailey, V. P., Bauman, B., Cardwell, A., Dillon, D., Gavel, D., Hartung, M., Hibon, P., Perrin, M. D., Rantakyrö, F. T., Sivaramakrishnan, A., and Wang, J. J. (2016).
Performance of the Gemini Planet Imager's adaptive optics system.
Applied Optics, 55(2):323–340.

Tallis, M., Bailey, V. P., Macintosh, B., Hayward, T. L., Chilcote, J. K., Ruffio, J.-B., Poyneer, L. A., Savransky, D., Wang, J. J., and GPIES Team (2018).
Air, telescope, and instrument temperature effects on the Gemini Planet Imager's image quality.
In *American Astronomical Society Meeting Abstracts #231*, volume 231 of *American Astronomical Society Meeting Abstracts*, page 361.18.

Wang, J. J., Perrin, M. D., Savransky, D., Arriaga, P., Chilcote, J. K., De Rosa, R. J., Millar-Blanchaer, M. A., Marois, C., Rameau, J., Wolff, S. G., Shapiro, J., et al. (2018).
Automated data processing architecture for the gemini planet imager exoplanet survey.
Journal of Astronomical Telescopes, Instruments, and Systems, 4(1):018002.