QUANTIFYING THE OSCILLATORY EVOLUTION OF SIMULATED BOUNDARY-LAYER CLOUD FIELDS USING GAUSSIAN PROCESSES

GUNHO (LOREN) OH* AND PHILIP H. AUSTIN
DEPARTMENT OF EARTH, OCEAN, AND ATMOSPHERIC SCIENCES
UNIVERSITY OF BRITISH COLUMBIA

1 Introduction

Oscillatory motions in clouds have long been observed in cloud field simulations and observations at multiple scales. Cumulus clouds can be seen as a series of pulsating plumes with a period of 10-15 minutes (Heus et al., 2009; Malkus, 1952; Zhao and Austin, 2005a,b), which plays an important role in understanding moist convection. On a much larger scale, Feingold et al. (2017) have observed that the simulated cloud size distribution oscillates over time with spectral power at a period of approximately 80 minutes. Specifically, they found in their simulations that xthe cloud field alternates between a relative abundance of large clouds breaking up into smaller plumes which then reform into larger clouds.

Gaussian process regression provides an alternate way to analyze periodic time variations of noisy processes. Here, we will apply Gaussian process regression to a large eddy simulation of marine boundary layer clouds similar to that modeled in Feingold et al. (2017). In Section 2 we define the time-series used to represent the changing cloud-size distribution and briefly discuss how Gaussian process (GP) regression is used to find periodic structure. In Section 3 we use GP regression and a more traditional spectral analysis to show that the large eddy simulation exhibits periodicity at both 45 minute and 78 minute timescales.

The original poster and all the code used to produce both the poster and the figures in this abstract are available for download from github at `https://github.com/phaustin/gaussian_processes_ams_2018.git`

loh@eoas.ubc.ca.

2 Methods

2.1 Model Description

To obtain cloud field statistics, we analyze the results from a high-resolution, large-eddy simulation (LES) using the System for Atmospheric Modelling (SAM; Khairoutdinov and Randall, 2003) based on the Barbados Oceanographic and Meteorology Experiment (BOMEX) case. the LES model run has been performed with a grid spacing of 25 m, a time step of 1 second, over a 13 km $\times$ 13 km $\times$ 3.2 km domain. The model run includes a two-moment microphysics scheme developed by Morrison et al. (2005a,b). We have also performed a number of boundary-layer simulations as well, and found that the results are more or less consistent. For the sake of brevity, therefore, we will focus on the analysis of the BOMEX case at the moment.

2.2 Cloud Size Distribution

Studies have shown that satellite observations Benner and Curry (1998); Koren and Feingold (2011); Zhao and Di Girolamo (2007) as well as model simulations Dawe and Austin (2012); Heus and Seifert (2013); Jiang et al. (2008); Neggers

et al. (2003) confirm that the cloud size distribution can be modelled by a negative-power law distribution as a function cloud cloud area a. That is,

$$\mathcal{P}(a) = Aa^{-b} \tag{1}$$

where $\mathcal{P}(a)$ is the frequency of clouds appearing in each cloud size bin (between a and $a + \mathrm{d}a$ in units of m^{-2}, and A and b are coefficients describing the characteristics of the cloud size distribution.

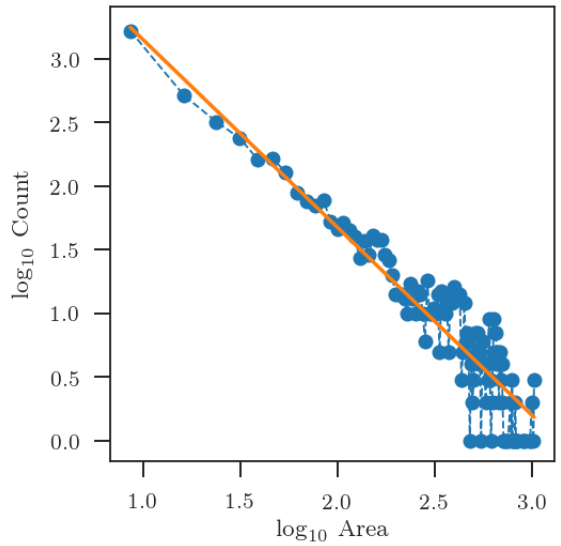

FIGURE 1. A log-log plot of a sample cloud size distribution $\mathcal{P}(a)$ as a function of cloud area a. The distribution in blue denotes the frequency of cloud samples in each bin between a and $a + \mathrm{d}a$. The result of the linear ridge regression is shown as the orange line.

We then take the log of Equation (1) to obtain

$$\log \mathcal{P}(a) = \log A - b \log a \tag{2}$$

which allows us to study the linear relationship between the cloud size distribution and cloud size in terms of the *slope* b and *intercept* A. Unless otherwise noted, we will only refer to the *slope* b as the main characteristic of the cloud size distribution $\mathcal{P}(a)$.

Figure 1 shows the cloud size distribution $\log \mathcal{P}(a)$ of the simulated cloud field 4 hours into our BOMEX simulation. As shown, the distribution is well described by a linear curve, obtained by performing a linear ridge regression yielding $A = 4.62$ and $b = -1.47$.

We can then construct the cloud size distribution time series by repeating the calculation for the slope b for the entire simulation period. That is, we accumulate the result of the ridge regression every minute for the entire duration of our BOMEX simulation. The exact value of A and b depends on a number of factors, such as the bin size and the definition of the *cloud area*, but it does not strongly affect the oscillation of these parameters.

As we can see in Figure 2, the resulting time series distribution of the slope parameter b is very noisy. It is not surprising, however, both because the underlying physical processes are stochastic, and because the calculation of the slope parameter involves uncertainties; for example, the choice of bin size as well as the regression method (Figure 1) could easily change both the slope and the intercept of the cloud size distribution.

There are a number of ways to remove long-term trends in a time series data, and here we simply calculate the mean distribution curve using a Bayesian ridge regression, representing the long-term changes in our data (Figure 3), and subtract the mean distribution from the time series data, resulting in a time series distribution without any long-term trends.

The resulting distribution (*c.f.* Figure 4) appears to be periodic, although it is very difficult to isolate a single wave-like element consistent over the entire duration of the time series. Still, it simplifies the Gaussian process regression as we can now assume that the prior distribution only needs to represent the oscillatory nature of the time series distribution of the slope parameter b (see Section 2.3 for more details).

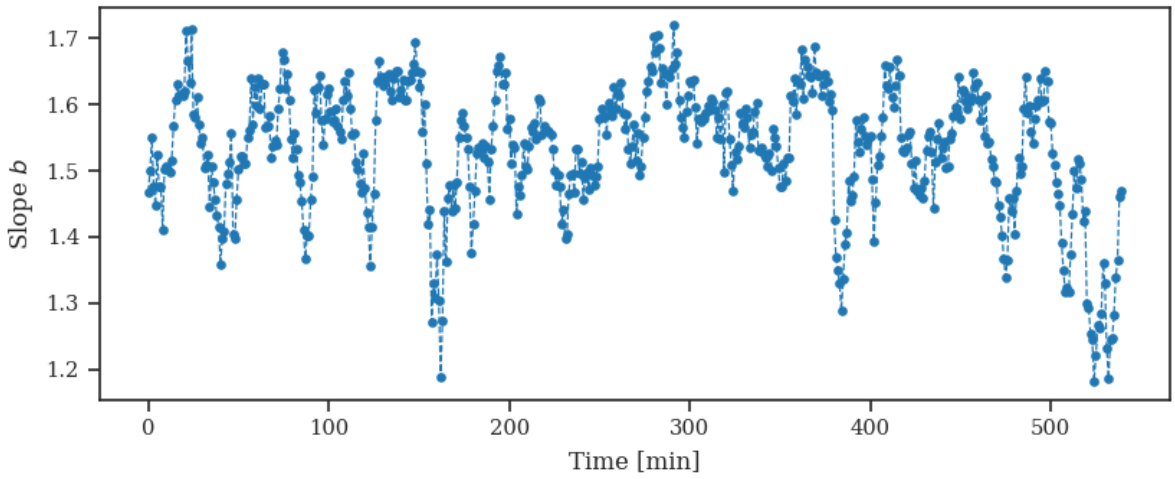

FIGURE 2. A time series distribution of the slope b of the linear fit of the cloud size distribution $\mathcal{P}(a)$ using ridge regression.

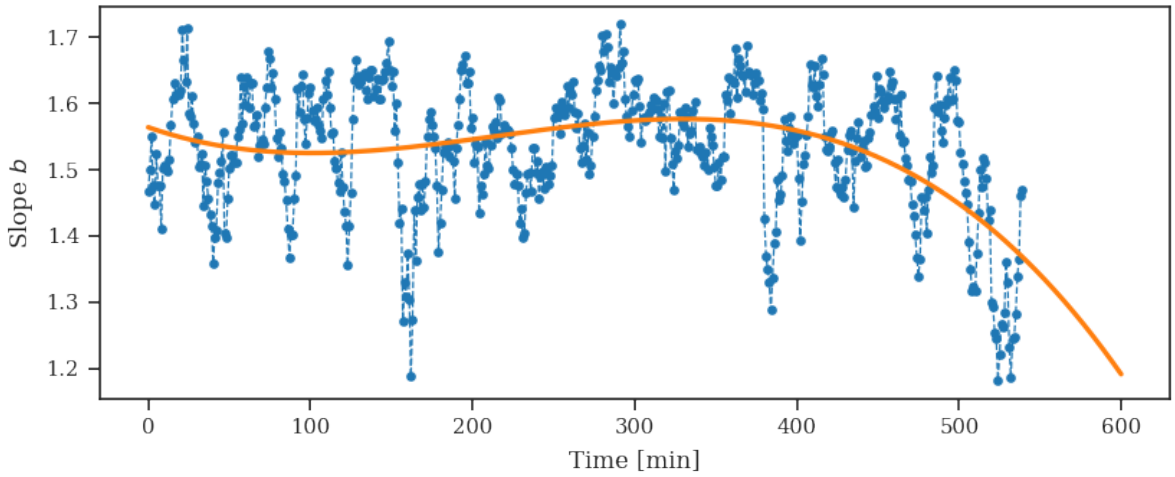

FIGURE 3. Same as Figure 2, but using Bayesian ridge regression to estimate the mean curve (orange line).

2.3 Gaussian Process Regression

Gaussian process regression is a Bayesian approach to modelling statistical distributions. A *Gaussian process* is a collection of random variables, any finite number of which have a joint Gaussian distribution. A Gaussian process is completely specified by its mean function and covariance function. The mean function $m(\boldsymbol{x})$ reflects the expected function value at input $\boldsymbol{x}$:

$$m(\boldsymbol{x}) = \mathbb{E}[f(\boldsymbol{x})],$$

that is, the average of all functions in the distribution evaluated at input $\boldsymbol{x}$. The prior mean is often set to $m(\boldsymbol{x}) = 0$. The covariance function $k(\boldsymbol{x}, \boldsymbol{x'})$ models the dependence between the function values at different input points $\boldsymbol{x}$ and $\boldsymbol{x'}$:

$$k(\boldsymbol{x}, \boldsymbol{x'}) = \mathbb{E}[(f(\boldsymbol{x}) - m(\boldsymbol{x}))(f(\boldsymbol{x'}) - m(\boldsymbol{x'}))]$$

and we can finally write the Gaussian process as

$$f(\boldsymbol{x}) \approx \mathcal{GP}(m(\boldsymbol{x}), k(\boldsymbol{x}, \boldsymbol{x'})).$$

The covariance between pairs of random variables is specified by the aptly named *covariance*

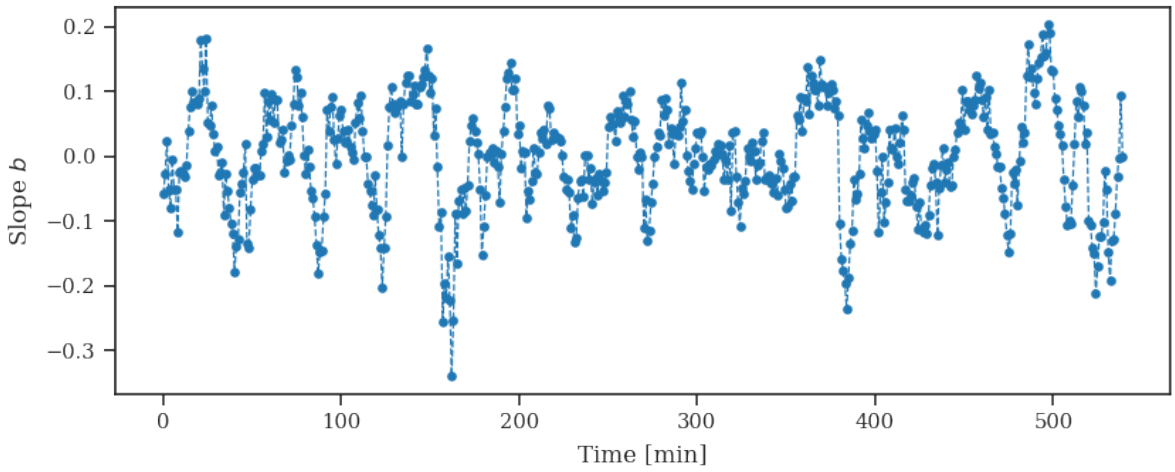

FIGURE 4. Result of de-trending the time series distribution in Figure 2.

function $k(\boldsymbol{x}, \boldsymbol{x'})$. The most popular choice of a kernel is the radial-basis function kernel, also called the squared exponential (SE) kernel,

$$k(x_i, x_j) = \exp\left(-\frac{1}{2}d\left(\frac{x_i}{l}, \frac{x_j}{l}\right)\right)$$

which is characterized by a length-scale parameter l.

For the purpose of periodicity detection, we are more interested in *periodic* covariance function

$$k(x_i, x_j) = \exp\left(-2\frac{\sin^2\left(\frac{\pi}{T}d(x_i, x_j)\right)}{l^2}\right)$$

which is specified by a length-scale parameter l and a periodicity T.

We can now sample the values of f at each set of input points $\boldsymbol{X}_*$ from the $\mathcal{GP}$ by sampling from a multivariate normal distribution

$$\boldsymbol{f}_* \sim \mathcal{N}(0, K(\boldsymbol{X}_*, \boldsymbol{X}_*))$$

where $\boldsymbol{f}_* = [f(\boldsymbol{x}_1^*), \ldots, f(\boldsymbol{x}_n^*)]^\top$, a sample from the distribution of functions evaluated at the corresponding input point.

Suppose we have made a series of noisy observations $y = f(\boldsymbol{x}) + \epsilon$ with noise ϵ. Assuming $\epsilon \approx \mathcal{N}(0, \sigma_n^2)$, the prior for these observations becomes

$$\text{cov}(\boldsymbol{y}) = K(X, X) + \delta_n^2 I$$

and the conditional distribution is then

$$\boldsymbol{f}_* \mid X_t, \boldsymbol{y_t}, X_* \approx \mathcal{N}(\overline{\boldsymbol{f}}_*, \text{cov}(\boldsymbol{f}_*)).$$

The posterior distribution is also a Gaussian process with mean

$$m(\boldsymbol{x}) = K(\boldsymbol{x}, \boldsymbol{X_t})[K(\boldsymbol{X_t}, \boldsymbol{X_t}) + \delta_n^2 I]^{-1}\boldsymbol{y_t}.$$

Therefore, having made the initial observations (on *training sets*), we can calculate the necessary terms to obtain posterior distribution and its kernel. The kernel is usually defined by a few hyper-parameters that are inferred from the data. For this reason, the bulk of GP regression method is to construct the kernel from possibly a number of covariance functions (in order to encode the prior assumptions about the observation), and obtaining the hyper-parameters from the observations.

Since this is often very challenging in a real-world scenario, the hyper-parameters are obtained by maximizing the (log) marginal likelihood. The log marginal likelihood is defined as

$$\log p(\boldsymbol{y} \mid \boldsymbol{X}) = -\frac{1}{2}\boldsymbol{y}^\top \boldsymbol{K}_y^{-1}\boldsymbol{y} - \frac{1}{2}\log|\boldsymbol{K}_y| - \frac{n}{2}\log 2\pi$$

where $K_y = K(\boldsymbol{X_t}, \boldsymbol{X_t}) + \sigma_n^2 I$. The first term measures how well the current kernel reproduces $\boldsymbol{y}$, the second term measures the complexity of the model, and the last term is a constant used for normalization.

For a more proper and detailed description of the Gaussian processes, refer to Rasmussen and Williams (2006).

3 Results

3.1 Periodicity Detection with Gaussian Process Regression

We can now apply the Gaussian process regression method described in Section 2.3 to the de-trended time series data in Figure 4. After a series of hyper-parameter optimizations, the hyper-parameters for the posterior distribution converge towards two periods at 78 minutes and 43.5 minutes. The resulting posterior distribution can be seen in Figure 5.

The mean posterior distribution from the Gaussian process regression is a good fit to the de-trended time series distribution of the slope parameter b as well as the intercept A (not shown). It is possible to perform a few more iterations of the hyper-parameter optimization, but improvements to the periodicity estimates are small compared to the complexity of the resulting posterior distributions (*c.f.* Rasmussen and Williams (2006)).

Feingold et al. (2017) report that based on a Fourier spectral analysis, two consistent peaks are observed at $\approx$ 80 minutes and $\approx$ 15 minutes. The former corresponds well to the mean posterior distribution seen in Figure 5 at roughly 78 minutes. We have modified the covariance function to include an additional periodic kernel and observed that it corresponds to a 15-minute period (not shown).

However, we also find a prominent periodicity at roughly 45 minutes. Given the discrepancy, it was necessary to verify that the Gaussian process regression method is producing posterior distributions that correspond well to the oscillatory motions seen in Feingold et al. (2017).

3.2 Regression with Missing Data

Lastly, Gaussian process regression is a very robust method that can be used with *sparse* data. In order to show the robustness of the method, we have repeated the regression for the cloud size distribution time series data with missing values. Figure 6 shows the resulting posterior distributions based on the modified time series data where some of the values are ignored entirely from the regression process. The sample dataset consists of 60-minute *sampling* periods, followed by 30-minutes of missing data.

As seen in Figure 6, the Gaussian process regression can accurately reproduce our time series data with missing data segments. The method yields period of 43.5 minutes and 78 minutes, and while it is not as accurate as with all the available data (*c.f.* Figure 5), the difference is inconsequential.

It is still possible to improve the accuracy of the mean posterior distribution, but it should be noted that these estimates are based on a dataset with *noisy* data. After all, attempts to reproduce the noisy data at the expense of computational resources and model complexity will inevitably model the inherent noise as well. If one is interested in extrapolating the results (e.g. for time series forecasts), this will result in a less accurate predictions.

3.3 AUTOPERIOD *Method*

The Gaussian process regression method provides a robust way to identify underlying oscillatory motions in the presence of uncertainties. However, it is necessary to confirm that the estimated periodicity values correspond well to those observed in Feingold et al. (2017). To this end, we implemented the method suggested by Vlachos et al. (2006), who use a combination of Fourier spectral analysis and autocorrelation function (ACF) to quantify the oscillatory motions in the observed time series data.

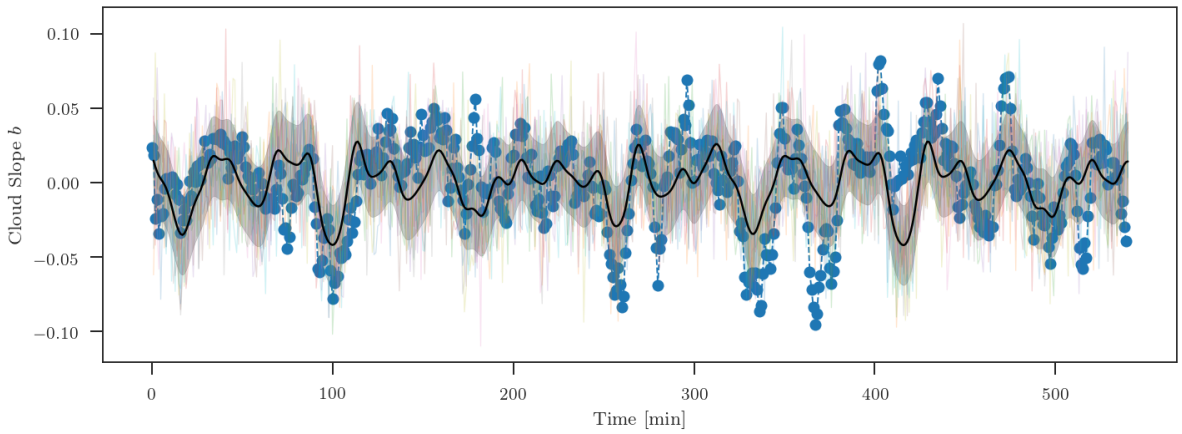

FIGURE 5. The time series distribution of the slope parameter b (blue), and the result of the Gaussian process regression, as a mean distribution of the posterior distributions (black). The shaded region represents one standard deviation from the average posterior distribution.

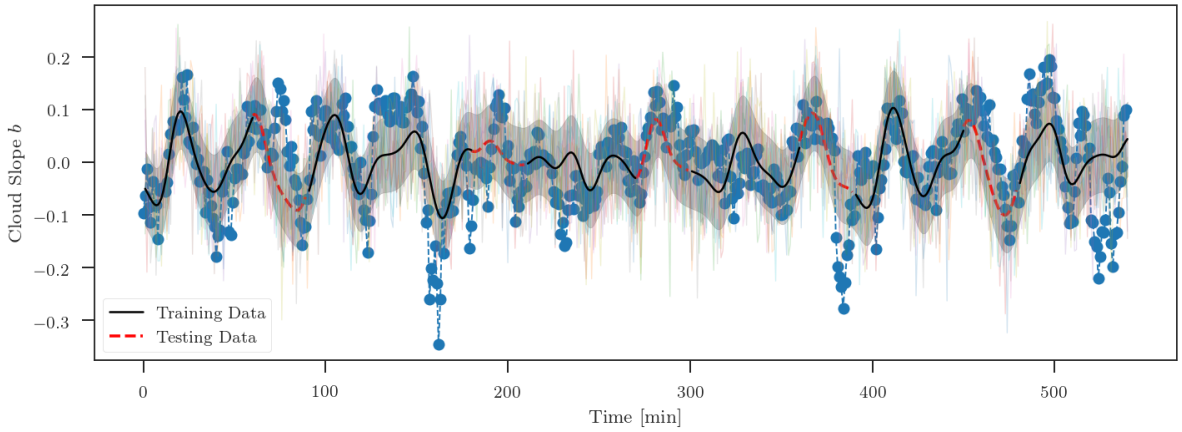

FIGURE 6. Same as Figure 5, but with missing data. The dashed red curve represents the mean posterior distribution where the data is missing from the training set used by the Gaussian process.

Normally, Fourier spectral analysis is employed to identify oscillatory motions in a noisy signal, as seen in Feingold et al. (2017) in this case. However, the estimated periods become rapidly unreliable for large periods. This is because the periodicity found in a periodogram (e.g. middle panel in Figure 7) represents a periodicity bin with a range of periods. Since the periods estimated in this example (Figure 5) are at least 15 minutes long, we determined that this method is rather insufficient to identify the periodic motions in the cloud size distribution. Also, the periodicity values that are not integer multiples of the bin width cannot accurately be determined by the Fourier spectral analysis.

On the other hand, as suggested by Vlachos et al. (2006), the ACF can be used to detect periodicities accurately even at large periods. But it is not sufficient by itself, mainly because the ACF distribution does not tell much about the significance of the peaks; for example, it is difficult to pinpoint the most significant extrema, and multiples of the estimated periods will appear as peaks as well.

We have applied the *AUTOPERIOD* approach (Vlachos et al., 2006) for periodicity detection using a more conventional technique, whose the results are shown in Figure 7. Interestingly, the most significant peak from the periodogram (middle panel) is at $T = 45$ minutes, which support our estimation using Gaussian process regression. We identified 8 most significant peaks from the periodogram, and located them on the ACF distribution (bottom panel). The most important periods are the ones at the two peaks, which correspond to $T = 45$ minutes and $T = 78$ minutes. Still, the prominent oscillating motion in the cloud size distribution appears to be at 45-minute period, as the 78-minute period is only the 6th significant candidate based on the Fourier spectral analysis.

4 Conclusion

The major challenges in identifying the periodic nature in cloud dynamics is twofold. On one hand, it is very difficult to isolate and quantify the oscillatory motions in the observed phenomena, from the oscillations in the cloud top heights (Heus et al., 2009) to those in the cloud size distribution (Feingold et al., 2017). On the other, even with sufficient data, periodicity detection is not an easy task in the presence of uncertainties.

We introduce the Gaussian process regression as a method to identify underlying oscillatory motions for noisy data, such as the slope parameter b of the cloud size distribution, motivated by the works of Feingold et al. (2017). We found that the Gaussian process regression is a robust periodicity detection method, even with noisy, sparse data (Section 3.2). For our BOMEX case, the most prominent periods were at 45 and 78 minutes, although the former appears to be more significant. We also verified this result using a more conventional method (Vlachos et al., 2006), whose results were consistent (albeit less accurate) with the Gaussian process regression.

There are a number of ways a Gaussian process regression can be used. For example, we have identified consistent oscillatory motions in the individually-tracked boundary-layer cloud top heights at periods of 14-17 minutes. The cloud top time series for boundary-layer cloud top height is effectively a large set of noisy, sparse data as the cumulus cloud cores are generally short-lived but oscillate at relatively long periods. We hope to be able to apply the Gaussian regression method to a number of oscillatory motions of clouds as observed over a multitude of scales.

Acknowledgments

This research was enabled in part by support provided by the Microsoft Azure for Research Program and Compute Canada.

See `https://github.com/phaustin/gaussian_processes_ams_2018.git` for a Jupyter notebook used to generate the abstract/poster figures.

References

Benner, T. C. and J. A. Curry, 1998: Characteristics of small tropical cumulus clouds and their impact on the environment. *J. Geophys. Res.*, **103 (D22)**, 28 753–28 767.

Dawe, J. T. and P. H. Austin, 2012: Statistical analysis of an LES shallow cumulus cloud ensemble using a cloud tracking algorithm. *Atmos. Chem. Phys.*, **12 (2)**, 1101–1119.

Feingold, G., J. Balsells, F. Glassmeier, T. Yamaguchi, J. Kazil, and A. McComiskey, 2017: Analysis of albedo versus cloud fraction relationships in liquid water clouds using heuristic models and large eddy simulation. *J. Geophys. Res. Atmos.*, **122 (13)**, 7086–7102.

Heus, T., H. J. J. Jonker, H. E. A. Van den Akker, E. J. Griffith, M. Koutek, and F. H. Post, 2009: A statistical approach to the life cycle analysis of cumulus clouds selected in a virtual reality environment. *J. Geophys. Res.*, **114 (D6)**, 97.

Heus, T. and A. Seifert, 2013: Automated tracking of shallow cumulus clouds in large domain, long duration large eddy simulations. *Geosci. Model Dev.*, **6 (4)**, 1261–1273.

Jiang, H., G. Feingold, H. H. Jonsson, M.-L. Lu, P. Y. Chuang, R. C. Flagan, and J. H. Seinfeld,

2008: Statistical comparison of properties of simulated and observed cumulus clouds in the vicinity of Houston during the Gulf of Mexico Atmospheric Composition and Climate Study (GoMACCS). *J. Geophys. Res.*, **113 (D13)**, 781.

Khairoutdinov, M. F. and D. A. Randall, 2003: Cloud Resolving Modeling of the ARM Summer 1997 IOP: Model Formulation, Results, Uncertainties, and Sensitivities. *J. Atmos. Sci.*, **60 (4)**, 607–625.

Koren, I. and G. Feingold, 2011: Aerosol–cloud–precipitation system as a predator-prey problem. *Proceedings of the National Academy of Sciences of the United States of America*, **108 (30)**, 12 227–12 232.

Malkus, J. S., 1952: The slopes of cumulus clouds in relation to external wind shear. *Q. J. R. Meteorol. Soc.*, **78 (338)**, 530–542.

Morrison, H., J. A. Curry, and V. I. Khvorostyanov, 2005a: A New Double-Moment Microphysics Parameterization for Application in Cloud and Climate Models. Part I: Description. *J. Atmos. Sci.*, **62 (6)**, 1665–1677.

Morrison, H., J. A. Curry, M. D. Shupe, and P. Zuidema, 2005b: A New Double-Moment Microphysics Parameterization for Application in Cloud and Climate Models. Part II: Single-Column Modeling of Arctic Clouds. *J. Atmos. Sci.*, **62 (6)**, 1678–1693.

Neggers, R. A. J., H. J. J. Jonker, and A. P. Siebesma, 2003: Size Statistics of Cumulus Cloud Populations in Large-Eddy Simulations. *J. Atmos. Sci.*, **60 (8)**, 1060–1074.

Rasmussen, C. E. and C. Williams, 2006: *Gaussian Processes for Machine Learning*. MIT Press.

Vlachos, M., P. S. Yu, V. Castelli, and C. Meek, 2006: Structural Periodic Measures for Time-Series Data. *Data Min. Knowl. Discov.*, **12 (1)**, 1–28.

Zhao, G. and L. Di Girolamo, 2007: Statistics on the macrophysical properties of trade wind cumuli over the tropical western Atlantic. *J. Geophys. Res.*, **112 (D10)**, 847.

Zhao, M. and P. H. Austin, 2005a: Life Cycle of Numerically Simulated Shallow Cumulus Clouds. Part I: Transport. *J. Atmos. Sci.*, **62 (5)**, 1269–1290.

Zhao, M. and P. H. Austin, 2005b: Life Cycle of Numerically Simulated Shallow Cumulus Clouds. Part II: Mixing Dynamics. *J. Atmos. Sci.*, **62 (5)**, 1291–1310.

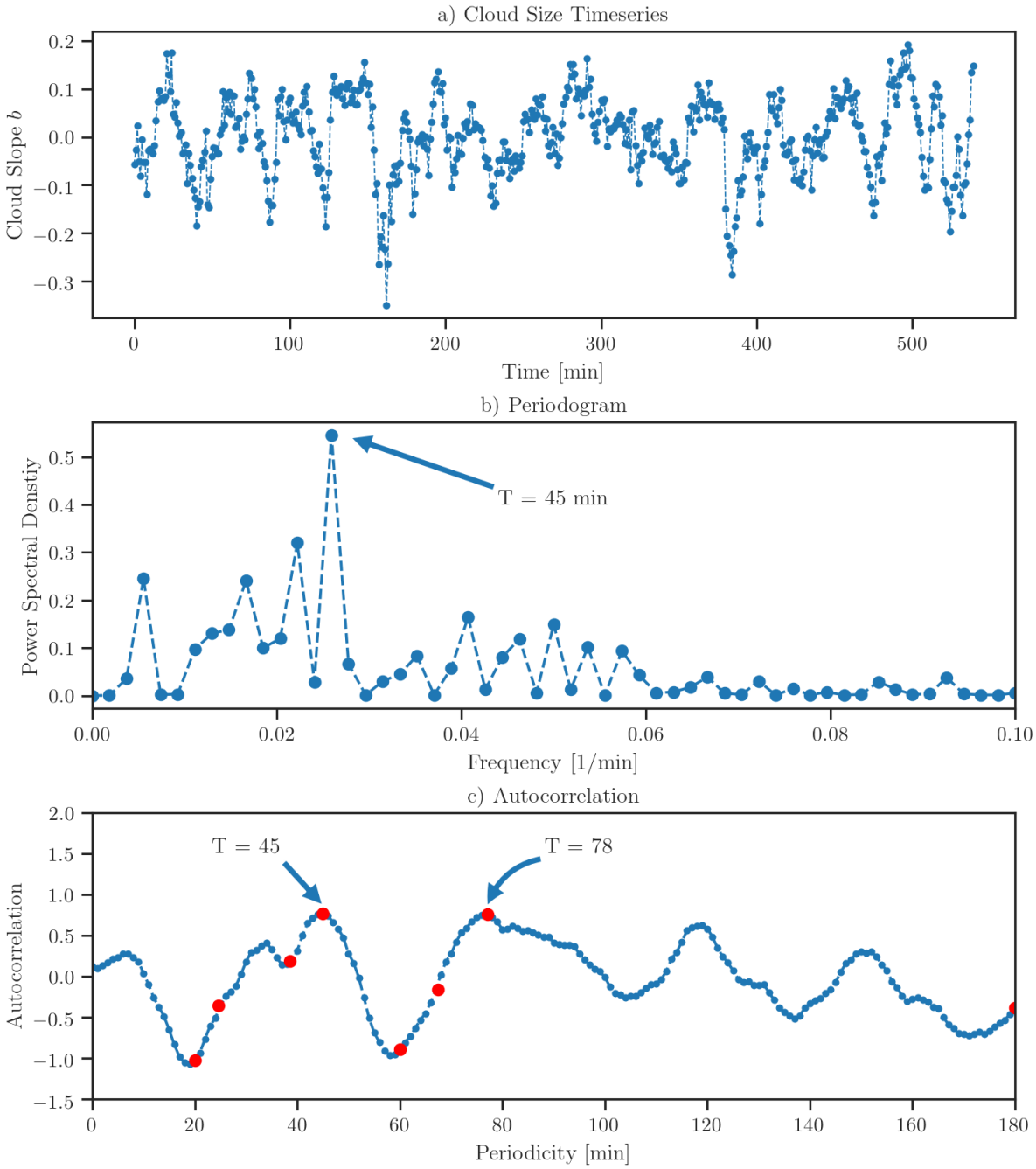

FIGURE 7. A more traditional approach to periodicity estimation using the spectral analysis based on Vlachos et al. (2006). a) The target time series distribution of the slope parameter b (Section 2.2). b) The periodogram of the time series distribution, given by the squared length of each Fourier coefficient. c) The (circular) autocorrelation function (ACF) of the time series distribution. The red dots represent 8 most prominent peaks on the Fourier spectrum in b).