Response to Comments from Reviewer 1

amt-2019-200
Total variation of atmospheric data: covariance minimization about objective functions to detect conditions of interest
Nicholas Hamilton

General comments: This paper offers a new interesting method to analyze a multivariable atmospheric data set. The method is clearly described, and the data analysis is thorough. It seems like a very versatile method with potentially many different usage scenarios. The manuscript would benefit by considering the points below. The language could be simplified at times and some sentences could be broken up for easier reading.

> *Thank you for taking the time to add your thoughts to the submitted manuscript. In addressing them, I think you will see that the manuscript has been greatly improved. Below, I have provided a brief response to each of the points you raised in the review of my work and, where appropriate, also included any additions or subtractions from the manuscript. In addition, I have edited the text in the manuscript to simplify the language where possible to increase readability. The manuscript has been greatly improved, due to your comments and the review process.*

Abstract: The problem statement in the abstract could be shortened, while at the same time there could be more information on the subject/method itself. The abstract would also be improved by including main results and findings.

> *Following the suggestion of the reviewer, the abstract has been revised to be more concise, while more clearly communicating the central contribution of the work. Given that the manuscript is focused on the introduction of a method, the abstract now points out the merits of the methods and points toward the sensitivity due to outliers as quantified through the Mahalanobis distance.*

Sections 2 and 3 could be shortened a bit. This would give the main part (section 4) of the paper more focus. A suggestion: Perhaps Table 1 and corresponding text could be removed as it is not so relevant for the focus in the paper.

> ***Sections 2 and 3*** *detail common steps used in the quality control of the atmospheric data and the aggregate statistical methods for wind energy. I feel that these sections are necessary to properly establish the narrative of the manuscript and to differentiate the total variation method introduced in the paper. As suggested by the reviewer, these sections have been revised where possible to make them more compact, while keeping their content clear and concise.*

It is not completely clear what parts of the method are novel and what has been done before. This could be pointed out.

> *The method relies on the classical understanding of correlated signals common to the analysis of physical systems. There are parallels with previous work as noted in the literature portion of the introduction. To reinforce this in the work, new references have been added for generalized variance, and a statement has been added to distinguish the novel contribution of the method developed in the paper*

and its application.

"The total variation, V, of a given regularized data block, D, is expressed as the determinant of the respective correlation matrix,

V=det(C) (6)

Larger values ofVindicate that the data points are more dispersed in the condition space. In the observational data of the atmosphere discussed here, V>0. The case of V= 0 would indicate that the full n−dimensional condition space is not occupied and some of the variables are perfectly correlated with, i.e. linearly dependent on, some of the others. Metrics of the variation of a multivariable dataset have some history in the literature. Notable past contributions include the pooled10variance method to estimate population variance from those of distinct samples Ruxton (2006), and the 'total' or 'overall' variability Goodman (1968); Anderson (1962) which combine variances of individual variables either linearly or in a sum of squares sense. The generalized variance (Wilks, 1932; Sengupta, 2004), shares a common formulation withV, but has historically been applied to a p−dimensional random vector. In contrast, the total variation merges n distinct variables, whose relationship need not be known a priori, and seeks the determinant of the associated correlation matrix"

The paper would benefit from a stronger discussion, perhaps in a dedicated section of its own.

Because the manuscript is focused mainly on the development of a method, rather than on analysis of a physical system, I feel that a Discussion section would not add clarity to the work, but rather would obfuscate the merits of the method with details about a single application. Instead, ***the discussion of benefits and potential drawbacks of the method have been expanded****.*

A thought: When we apply specific objective functions, we generally decrease the total variation and find conditions of interest by minimizing the total variance. Would we find similar conditions by not applying any objective functions and maximizing the total variation instead?

The inverse approach is not expected to identify conditions of interest, as shown in Figure 8(b). Maximizing the total variation is not guaranteed to identify any specific condition, but rather identify those that agree with the objective functions the least. For example, instead of applying a linear objective function to the data block to find the cleanest wind speed ramps (as in Figure 10) and selecting the time periods with minimal V is not the same as choosing the maximum V without the application of objective functions.

Are there any available codes or scripts with this method implemented?

No codes or demonstration scripts have been included as part of this submission. Given that the method is relatively straightforward, involving only a handful of well-defined mathematical operations, it seems unnecessary to provide a template for applying the method.

Specific comments:

In the abstract lines 3-4: “Most often, conditions of interest are determined as those that occur most frequently. . .” And similarly, p. 3, lines 12-13: “Within any wind plant data. . .” This statement would benefit from a reference, because it could be argued that the opposite often holds true. E.g. for wind turbine site assessment and certification, the conditions of interest are critical weather and extreme conditions.

*The reviewer is correct. Essentially, conditions of interest are necessarily defined by the research, and often times may be focused on infrequent or extreme conditions as these have particular relevance to wind plant behavior and operations. T**he phrasing has been changed** to emphasize that in validation of numerical models, commonly occurring conditions are often selected for comparison as they provide the most converged statistics with the least uncertainty due to sample size.*

Abstract, **“Atmospheric conditions relevant for wind energy research include stationary conditions, given the need for well-converged statistics for model validation, as well as conditions observed less frequently, such as extreme atmospheric events, which are used in wind turbine and wind plant design.“**

P. 3, **“Within any wind plant data, conditions of value for validation are typically identified by way of aggregate statistical metrics or by identifying “well-behaved” time periods exhibiting a dynamical event or atmospheric condition of interest.“**

Introduction, p. 2, lines 25-27: This is quite a strong statement – it would benefit from a citation or further argumentation. Introduction, p. 2, lines 27-29: Direct comparison of statistical quantities to what? Why does that discount the coupling between quantities that underpin atmospheric physics? This could be clarified and explained further.

In order to clarify the sentence, the text has been modified and citations have been added to support the statement that,
“Consideration of these variables independently may not provide a complete picture of the state of the atmosphere, as they are inherently correlated (Holtslag and Nieuwstadt, 1986; Kaimal et al., 1976); each variable offers a limited range of insights as to the dynamical state of the atmosphere relevant to the operation of wind energy assets. Further, and perhaps most importantly, consideration of statistical quantities (measures of central tendency, variability, or higher statistical moments) may discount the inherent coupling between quantities of interest that underpin atmospheric physics (Hannesdóttir and Kelly, 2019; Preston et al., 2009; Shahabi and Yan, 2003).”

Figure 1 a): The numbers on the colorbar are missing the number 9 in front of 00 and 50.

It is not entirely clear to me what is in error for Figure 1(a). The colorbar appears to be scaled correctly and the labels appropriate. A new figure has been placed in the updated version of the manuscript, but it may be that the pdf rendering through the journal website or pdf viewer may have created some display error.

Equation 4-5: Should just be a single equation with one number. Further, I cannot see how the matrix multiplication would result in the covariance matrix. Unless the average of each column has been subtracted from the values in D Ì´C and the values have been divided by m. If that is the case, it should be mentioned. At this point in the paper normalization has not been mentioned.

Equations 4 and 5 have been combined in the manuscript and a factor of 1/(m-1) has been added for completeness sake. As noted by the reviewer the mean of each channel is removed during the data standardization step, otherwise the correlation matrix would not follow the traditional formulation. The statement about normalization of the data has been moved from Section 4.1 up to the definition of D, where it is more appropriate.

Figure 6: It does not seem that the histogram adds up to 100%. Has the data been cut off at Total Variation=0.3? It would be better to show the whole range of the Total Variation.

The reviewer is correct, and the upper tail of the distribution has been truncated. These distributions include some very high values of V, and have been truncated to emphasize the lower values, where differences between the two distributions are most visible. A note has been added to the caption of Figure 6 clarifying this point, ***"Both distributions in Fig. 6 have been limited to V≤0.30 to emphasize differences between the two data block lengths. In either case, the distribution is positively skewed, and high values of V exist with very low frequency."***

Figure 7 a) is not mentioned anywhere in the text. It should be commented and explained in the text.

Thank you for pointing out this oversight. ***Figure 7(a) is now referenced*** *in the paragraph immediately preceding it where the text is focused around the distribution of observations in the condition space that correspond to the minimum and maximum values of V.*

"Fig. 7(a) shows that the periods with minimal values of V have time series that appear constant and experience only small stochastic variations within each channel and that periods with large values of V exhibit more spread"

Page 12, equation 12-13. Again, should just be one equation. Also, what objective function is used for the TI? It is not mentioned.

Equations (12) and (13) have been combined as suggested by the reviewer. The objective function blocks have also been clarified in equations (8)-(10), showing explicitly that in each case, functions are 0 unless otherwise specified.

Page 12: When the objective function eq. 9 is applied to the wind speed, what objective functions are then applied to the direction change and TI at the same time?

In each of the demonstrated regularization schemes, the listed objective function is applied to the specified data channel and the others remain unaltered. That is the other objective functions remain zero. The equations have been modified to highlight the objective function blocks used for regularization, rather than specifying the functions alone. This should make the regularization schemes more clear to the reader.

Page 12, lines 17-18: "Defining specific functions, even of the same forms, would likely increase the average value and spread of V. . .". Are you certain of this? According to Figure 9 a) the average value and spread of V has decreased by subtracting the objective functions from the data. As you mentioned, subtracting the objective function acts as detrending, and therefore it should be expected that the total variation would

always decrease, as it is only the stochastic part of the data that determines the covariance of the remaining data.

The reviewer is correct, general detrending the data should reduce the resultant value of V. The intent of this statement was to convey the idea that if you remove the wrong trend from a time period, you may inadvertently increase V. This is not expected to be the case when using the least-squares minimization to determine fit coefficients as in the article. When the coefficients are prescribed a priori, there is no guarantee that the covariance would be reduced by removing the objective function. The offending sentence has been edited to read,
"Defining the coefficient values ahead of time would likely increase the average value and spread of V; for example, it is not expected that a wind speed ramp with specific slope and vertical offset would fit every time period well, and thus would not necessarily reduce the total variation for that period."

Figure 9: Includes two subfigures named (d). Also, these are not mentioned in the text, but should be. What is fit frequency - is it connected to eq. 11? Could you elaborate?

Thank you for pointing this out. The journal prefers subfigures to be collected into single image files, and this was overlooked. The fit frequency refers to the coefficient c_0 from equation 10. **The captions in Figure 9 have been updated** to more clearly communicate what information is shown in the distributions.

Section 5: The data used in this section is synthetic, and provides a very illustrative example of the sensitivity. However, I wonder if the removed points can be interpreted as outliers. Could we not say that these are extremes? Maybe the outliers could be assigned standard deviations outside of the range [0, 10], to ensure that they represent "real" outliers due to e.g. measurement errors.

While it is certainly possible for extreme values to excluded as outliers, it should be considered which time periods will be identified as favorable via total variation. If no objective functions are supplied, the method is tuned to quantify the variability of the data about stationary conditions. Extreme values occurring during a given period will probably increase the respective value of V, but these periods should probably not be considered as stationary in any case. If the intent of quantifying V is to identify conditions that include extreme events (gusts, turbulent structures, weather fronts, etc.) the objective functions should be defined to highlight them. Use of the Mahalanobis distance assumes in the current work assumes that each variable is normally distributed within a given time frame. Accordingly, a Mahalanobis *distance of 3 implies that there is approximately 1.1% probability of a point being an outlier for two degrees of freedom (as in the outlier sensitivity study) and 2.9% for three degrees of freedom (as in the demonstration with atmospheric variable data). The particular value of the Mahalanobis distance threshold used, should take into account the number of degrees of freedom (i.e. the number of variables) considered in the data. A note to that effect has been added in Section 5.*
"Any point withχ >3is flagged as an outlier and eliminated. With two degrees of freedom (variables in the data block), values ofχ >3are expected to be observed with a probability of approximately 1.1% (Penny, 1996; Ben-Gal, 2005; Gellert et al., 2012)."

Page 16, line 25: ". . . the method is independent of the length of the data record. . .". How can this statement be supported by the current analysis?

This statement is intended to communicate that the method does not explicitly require a record of a particular length or resolution. The sentence has been revised to read,

"In addition, the method should be equally applicable to any data, regardless of which variables are part of the data block and for data of any length and resolution, provided that enough observations are present to ensure reasonably converged statistics."