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."

Response to Comments from Reviewer 2

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

There is some good content and work here, with a generalized method to find conditions of interest for multivariate timeseries; and (perhaps more importantly) inclusion of responsible application of a metric (Mahalonbis distance) to evaluate sensitivity of the method to outliers.

> *Thank you for taking the time to review my submission. I appreciate your concise and direct comments and, in addressing them, I think you will see that the manuscript has been greatly improved. It pleases me that the intended message of the work has been clearly understood and well received. 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.*

The title is perhaps not quite appropriate; "Total variation of atmospheric data" is rather vague and somewhat grandiose, not accurately capturing the essence of the work and connoting more results/applicability than demonstrated.

> *I think that your suggestion is correct. The title never felt like it was perfectly suited to the content of the manuscript. Accordingly, the title has been changed to, "**Atmospheric condition identification in multivariate data through a metric for total variation**", which I believe more concisely conveys the intent of the work and communicates its scope as the development of an analysis and quality control method.*

Some significant items of note, as a list:

In the abstract, 'periods' of interest is better expressed as 'conditions', both for the sake of validation and for getting conditional statistics (and towards making fair comparisons of statistics given some conditions).

> *I think that the suggested change from 'periods' to 'conditions' is appropriate. While the method is designed to quantify the total variability within a continuous time period, it is the identification of atmospheric events or conditions of interest that is the real objective.*

Stationarity and conditional statistics underpin this written work; these concepts should be integrated (and referenced, as found in various texts for atmospheric flows), at least starting with the literature review.

> *The reviewer is correct to point out that the concept of statistical stationarity is one of the main concepts driving the current work. From the fundamental turbulence perspective, the term stationarity is not really expected to apply to data from an*

inherently dynamical system (the atmosphere) over periods of this duration. However, the term 'stationary' is also familiar to the atmospheric science community, and has now been mentioned explicitly, as suggested by the reviewer. A statement has been added to Section 4 to underpin the importance of stationarity, ***"Statistical stationarity (i.e. time-independence of statistical quantities) is a common consideration in turbulence and atmospheric science (Chenge and Brutsaert, 2005; Metzger et al., 2007; Vincent et al., 2010, 2011; Guala et al., 2011). Stationarity is not often assumed for wind energy research and modeling applications, although it is rarely quantified or even considered in validation data****."*

In your literature review, a key method/scheme for event detection (beyond wavelets) appears to be missing: i.e., reference-signal (or ideal signal) approaches based on Hilbert transform, as in Hristov et al (1998, PRL 81 no.23), used in various literature (e.g. Kelly, Wyngaard & Sullivan 2009).

I would like to thank the reviewer for pointing out this method for detection of atmospheric conditions. A statement has been added to the introduction including the above references.

"Another method for parsing atmospheric conditions found in the literature leverages the Hilbert transform, which convolves time series signals with the Cauchy kernel and results in a phase-shifted set of Fourier components. This method has been used successfully to relate ocean wave conditions to atmospheric conditions through the use of a reference signal (Hristov et al., 1998) and has successfully been extended to turbulence modeling (Sullivan et al., 2000; Kelly et al., 2009) and to relate turbulent motions of various scales within the atmospheric boundary layer (Mathis et al., 2009). Previous use of the reference-signal method (Kelly et al., 2009) required the use of a periodic reference signal, which does not lend itself easily to the detection of non-periodic atmospheric events, and strongly-correlated ocean wave and turbulent velocity data, which are not available for the majority of wind plant data sets.*"*

When you mention "direct comparison of statistical quantities", it appears that you are trying to refer to statistics based on marginal distributions (or marginal statistics), are you not? In statistical parlance, one contrasts between marginal and conditional statistics.

The reviewer is correct, and that sentence was intended to describe comparison of marginal statistical quantities. The sentence in the introduction has been changed to read,

"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. Direct comparison of the marginal distributions of atmospheric variables aggregates observations without regard to the value of other, potentially correlated variables. Even the use of conditional statistical distributions or measures discounts any dynamic coupling between them and may not fully describe the nature of the atmospheric physics (Hannesdóttir and Kelly, 2019; Preston et al.,

2009;Shahabi and Yan, 2003)."

The premise "In lieu of a time series of Richardson number or the Monin-Obukhov stability parameter, turbulence intensity (TI) is used in the current demonstration as a proxy for stability" is fundamentally problematic. That is, the balance of mechanical (shear) production, buoyant production or destruction, and dissipation ε (defining the 'simple' conditions where Monin-Obukhov similarity applies) results in TI being a proxy for stability only for flows/conditions with the same dissipation rate (Kelly, Larsen, Dimitrov & Natarajan, 2014). So your results per TI are conditional on ε, and do not act as such a proxy unless further constrained (e.g. via U assuming surface-layer similarity for ε.) Since stability is not really used in the paper, I suggest that you simply keep TI, and change the justification for its use: σu and TI are important for driving turbine loads (e.g. Dimitrov, Kelly, Vignaroli & Berg 2018).

Thank you for your concise description of the issue of regarding TI as a proxy for metrics of atmospheric stability. This is an important point to consider when making decisions as to how one should quantify the state of the atmosphere considering the data available. In the current case, as noted by the reviewer, stability is not discussed outside of the referenced section, given that temperature and/or heat flux information are not available for the data used in the current demonstration, it would probably be better to focus the narrative around TI as a relevant quantity of interest for wind turbine loads and wake modeling. The previous framing of the discussion arose from the intent to state that stability is an important factor in describing the state of the atmosphere, while conceding that TI is the quantity considered in many wind energy applications. The relevant excerpt has been changed to read,

"Data used in the current work does not contain any observations of the temperature or heat flux between the atmosphere and the ocean surface, and thus no estimate for the traditional stability metrics are available. Turbulence intensity (TI), although an imperfect proxy of atmospheric stability from a fluid mechanical or atmospheric perspective, provides some sense of the energy contained in the fluctuating flow field, and is well-suited for presenting the utility of the total variation method below. Additionally, TI is a quantity frequently used in the wind energy community to characterize wind plant operating conditions and structural loading of wind turbines (Kelly et al., 2014; Dimitrov et al., 2018) and is often accessible through instrumentation on met masts or wind turbine nacelles making it an appropriate choice for the current demonstration."

In section 3, where you write "without explicitly considering the evolution of atmospheric variables" you should mention stationarity as well. In the atmospheric sciences and boundary-layer meteorology this is typically considered, whereas it is often neglected in wind energy applications.

A similar point from the reviewer regarding the discussion of statistical stationarity has been addressed above. A brief statement has been added to Section 3, noted by the reviewer, reading,

"Considering atmospheric variables in terms of either their marginal distributions (as in Fig. 2 or their conditional distributions (as in Figs. 3 and 5) falls short of saying anything about the dynamics embedded in those observations. Steady-state wake models are defined to represent the time-

averaged flow behind a wind turbine and higher-fidelity models assume that the bulk flow speed and direction do not change in time. Effective validation of numerical modeling tools for wind energy requires that observations conform to stationary atmospheric flow (Chenge and Brutsaert, 2005; Metzger et al., 2007; Vincentet al., 2010, 2011; Guala et al., 2011) or represent a dynamic event of interest."

Figure 5: missing axis values/scales

I must apologize for the rendering of the figure. I believe that the axis labels were not included in the typeset document for some reason. In the revised version of the manuscript, Section 3, describing the statistical view of atmospheric conditions, has been reduced in length. Because the 3D histogram did not add significantly to the discussion of the distributions of atmospheric variables beyond the 2D histograms, the figure and associated discussion has been removed.

Section 4: can you interpret the total variation in terms of the multivariate components, to avoid obfuscation? Section 4.0 (p.8) is essentially taken from PCA; you should include reference to appropriate PCA text(s) and try to explain V for the reader. E.g., for readers not as 'fluent' in statistics, if the PC's (P) are orthogonal, then how are the covariances accounted for?

The formulation leading to the total variation does include an eigendecomposition of the covariance matrix and is in fact related derived from PCA. The method was defined this way because PCA was one of the methods originally considered during the analysis. Because the principal components are not identical to the original variances, they must include information from the covariances. That said, the sum of the principal components is also equal to the trace of the covariance matrix, which remains difficult to relate to the covariances between variables. In subsequent work, I found that the determinant of the covariance matrix also reduces the covariance matrix to a single metric that quantifies its variability. In fact, for the current study, the determinant method and the PCA method rank the variability of continuous time periods in the same order, although the numerical value is a bit different. The formulation has been updated using the determinant method, which also happens to be a more direct means at arriving at $\mathcal{V} $.

"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"

Is your V different than the 'overall' or 'total' variability found in literature?
It could help also to point out the difference between summative variance and V.

These are good points and, given their similarity, I have decided to answer together. I take it that the reviewer is suggesting that the total variation method be more clearly related or disambiguated from other statistical measures of variability. The metrics total variability, overall variability, and summative variance in common use have slightly definitions and interpretations from the total variation introduced in the current work. Briefly,

***Total variability** is defined as the sum of squares total of difference between expected or mean value and observed qualities.*

***Overall variability** refers generally to the variance or standard deviation of a population (i.e. a group of samples considered together).*

***Summative** or **pooled variance** refers to the inferred variance of a population of observations from the collection of sample variances.*

In contrast, the total variation used in the current work reduces the covariance between normalized variables to a single value through the determinant of the covariance matrix.

A close analog to this method is the generalized variance of a multi-dimensional random vector. Generalized variance was introduced by Wilks as a scalar measure of overall multidimensional scatter. However, in most formulations of generalized variance, the data are considered as a p-dimensional vector. The current work uses the same mathematical operations but applies them to distinct variables that have been merged into a matrix. Mechanically, the same operations are being applied to the data, but given the distinction in formulation, I have elected to maintain the current jargon of 'total variability'. A statement has been added to the introduction with references to some other metrics of variability.

"The metric used to quantify the overall variability of the atmosphere within any given time period is closely related to the generalized variance as per Wilks (1932); Sengupta (2004), but is distinct in that it is applied to a collection of variables rather than a multi-dimensional vector."

Figure 8: suggestion: use logarithmic scale on y-axis to compare more sensibly

I thank the reviewer for the suggestion, although I'm not sure I entirely understand what the purpose of logarithmic scaling would be. The figure displays the atmospheric variables considered during time periods with minimum or maximum values of V Given that the data do not span multiple orders of magnitude, rescaling the axes is not expected to add to the interpretation of the data.

Fig.9c: which "dimensionless slope" are you using here?

The dimensionless slope referenced in the caption of Figure 9c refers to the coefficient c_0 in eq. (7). While all of the coefficients in relationships seen in eqs. (7) - (9) are dimensionless due to the normalization of the variables, the phrasing is a bit difficult to follow. All of the subplots captions have been updated accordingly.

Fig.11: captions are swapped between (c) and (d).

Thanks for catching this oversight. The figure captions have been updated.

Please also note the supplement to this comment:

Additional (minor) comments found in the marked-up document have all been addressed in the manuscript. Thank you for the detailed review of the work. I feel that it is substantially improved due to your thoughtful comments.

Atmospheric condition identification in multivariate data through a metric for total variation

Nicholas Hamilton

National Renewable Energy Laboratory, Golden, Colorado, USA

Correspondence: Nicholas Hamilton (nicholas.hamilton@nrel.gov)

Abstract. Identification of atmospheric conditions within a multivariable atmospheric data set is a necessary step in the validation of emerging and existing high-fidelity models used to simulate wind plant flows and operation. 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. Aggregation of observations without regard to covariance between time series discounts the dynamical nature of the atmosphere and is not sufficiently representative of atmospheric conditions. Identification and characterization of continuous time periods with atmospheric conditions that have a high value for analysis or simulation sets the stage for more advanced model validation and the development of real-time control and operational strategies. The current work explores a single metric for variation of a multivariate data sample that quantifies variability within each channel as well as covariance between channels. The *total variation* is used to identify conditions of interest that conform to desired objective functions, such as stationary conditions, ramps or waves of wind speed, and changes in wind direction. The direct detection and classification of events or conditions of interest within atmospheric data sets is vital to developing our understanding of wind plant response and to the formulation of forecasting and control models.

1 Introduction

Parsing multivariate data sets that are ever growing in size and complexity can be a daunting task for researchers seeking to identify periods or events of interest in time series data (Preston et al., 2009; Shahabi and Yan, 2003). This is especially true for wind energy research seeking to validate high-fidelity numerical models against field observations (Barthelmie et al., 2015; Larsen et al., 2013; Sørensen and Shen, 2002). Wind plants operate continuously over time periods spanning years and across a broad range of atmospheric conditions, each of which implicitly impact the operation of the wind plant, either in terms of power production, operations and maintenance costs, or energy forecasting for grid integration.

Field observations of wind plants are typically collected by instrumentation mounted to wind turbines or meteorological towers, met masts, and by supervisory control and data acquisition (SCADA) systems. Wind plant data sets typically include measurements of wind speed and direction, local temperature and pressure, and wind turbine operational data, such as operational status, power production, and nacelle position. Each of the atmospheric quantities of interest may be classified as non-ergodic stochastic variables that are fundamentally connected (i.e. strongly interdependent).

Wind speed ramps are of particular interest in wind plant power forecasting due to the need to balance energy production against demand curves and in the planning of required reserves and base loads (Sevlian and Rajagopal, 2012; Zhang et al., 2014). Previous work has focused on forecasting of mesoscale changes in wind speed (Bossavy et al., 2013; Ferreira et al., 2011), generally concentrating on risk and reliability issues for wind turbines. Ramp event detection has been a research focus for more than a decade, (Cutler et al., 2007; Ferreira et al., 2013; Hannesdóttir and Kelly, 2019), and has produced some specific recommendations for individual turbine controls and the influence on operations and maintenance costs or activities. Previous research in wind speed ramps is not easily generalized to the identification and characterization of other dynamical events of interest, despite parallels in the detection process and considerations for wind turbine or plant operations and controls.

Detection of events in noisy data is of particular interest in the case of turbulent atmospheric data sets, especially given the need for more sophisticated forecasting systems (Belušić and Mahrt, 2012; Fulcher, 2018; Gamage and Hagelberg, 1993; Kang et al., 2014, 2017; Sun et al., 2015). One of the more common event detection methods leverages the continuous or discrete wavelet transform (Gamage and Hagelberg, 1993; Kumar and Foufoula-Georgiou, 1997; Lilly, 2017). Wavelet transforms leverage time-frequency signals designed to have specific properties that make them easy to use in signal processing applications. However, wavelet transformation remains computationally intensive and requires a fair amount of expertise to implement effectively and avoid the common pitfalls of signal shift sensitivity and the poor representation of phase and directionality (Taswell, 2001). A more direct method simply considers the covariance matrix of the input data, which represents the statistical spread of each data channel as well as cross-correlated variability (Eaton, 1983; Wasserman, 2013). Reducing the variability of a sample of multi-dimensional observations to a single metric is a necessary step to using numerical methods such as least-squares minimization for event detection and classification.

Another method for parsing atmospheric conditions found in the literature leverages the Hilbert transform, which convolves time series signals with a Cauchy kernel and results in a phase-shifted set of Fourier components. This method has been used successfully to relate ocean wave conditions to atmospheric conditions through the use of a reference signal (Hristov et al., 1998) and has successfully been extended to turbulence modeling (Kelly et al., 2009; Sullivan et al., 2000) and to relate turbulent motions of various scales within the atmospheric boundary layer (Mathis et al., 2009). Previous use of the reference-signal method (Kelly et al., 2009) required the use of a periodic reference signal, which does not lend itself easily to the detection of non-periodic atmospheric events, and strongly-correlated ocean wave and turbulent velocity data, which are not available for the majority of wind plant data sets.

Simultaneous observation of multiple thermodynamic and kinematic quantities reported by met masts are necessary to characterize the dynamical state of the atmosphere (Barthelmie et al., 2014; Hansen et al., 2012). Directly considering multiple disparate data channels simultaneously represents a challenge in that each quantity has different engineering units and that variation within each channel may occur over a distinct scale. Atmospheric conditions are frequently characterized by considering wind speed, wind direction, and turbulence intensity or thermal stability, each of which have different units, ranges, and statistical properties. 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. Direct comparison of the marginal distributions of atmospheric variables aggregates observations without regard to the value of other, potentially correlated variables. Even the use of conditional statistical distributions or measures discounts any dynamic coupling between them and may not fully describe the nature of the atmospheric physics (Hannesdóttir and Kelly, 2019; Preston et al., 2009; Shahabi and Yan, 2003).

The following work explores an application of numerical analysis methods to atmospheric data to identify continuous periods of interest within met mast time series data. The source of the data and their treatment are discussed briefly, although the wind plant and met mast are not in themselves imperative to the demonstration of the method or its utility. A discussion of aggregate statistical measures of the data is followed by a formal definition of the total variability of a block of time series data, and applications using the total variation as a metric to identify specific dynamical events of interest. The metric used to quantify the overall variability of the atmosphere within any given time period is closely related to the generalized variance as per Wilks (1932); Sengupta (2004), but is distinct in that it is applied to a collection of variables rather than a multi-dimensional vector. Finally, sensitivity of the method to outliers is analyzed, ending with a discussion of broader applications and extensions to the method.

2 Data and quality control

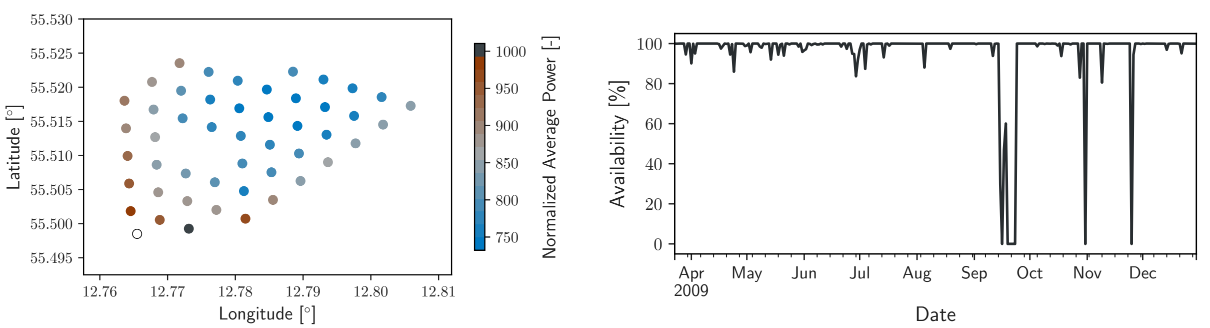
Data used to demonstrate the current method for detecting conditions of interest issue from met mast signals at the Lillgrund Wind Farm, located 10 km off the coast of southern Sweden in the Kattegat Strait. Lillgrund is comprised of 48 Siemens SWT-2.3-93 wind turbines and has a rated nameplate capacity of 110 MW. The layout of the Lillgrund wind plant is shown in Fig. 1(a), where each turbine location is denoted with a marker whose color is representative of the average power produced over the time period analyzed below. Operational data (SCADA, power production, turbine availability) from the wind farm are not discussed further in the following analysis, although a brief summary of future applications of the method is provided in the conclusions section, including thoughts on wind plant performance and SCADA data. Data used to demonstrate the calculation of total variation and identify periods of interest come from the met mast, located at the southwest corner of the wind plant, indicated in Fig. 1(a) with an open marker.

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. Kinematic and thermodynamic atmospheric quantities that are expected to have the greatest impact on the performance of a wind plant are the wind speed u, wind direction θ, and the atmospheric stability, considered either in an instantaneous or time-averaged sense. The stability of the atmosphere (typically quantified by the Monin–Obukhov stability parameter or the Richardson number) indicates the magnitude of buoyant production or destruction of turbulent kinetic energy (TKE) relative to shear production of TKE, and whether it represents either a source or sink of (vertical) momentum (Kumar et al., 2006; Wyngaard, 2010). Forcing in the momentum equations as indicated by the presence and sign of a buoyancy term is manifested in atmospheric flow as vertical turbulent mixing, and is an important overall factor in the energy balance relevant to wind plant operation. Thermal stability has a significant effect on atmospheric turbulence and the structure of wind turbine wakes, wake interaction, and thus the overall energy balance within the wind plant (Ali et al., 2019).

Data used in the current work does not contain any observations of the temperature or heat flux between the atmosphere and the ocean surface, and thus no estimate for the traditional stability metrics are available. Turbulence intensity (TI), although an imperfect proxy of atmospheric stability from a fluid mechanical or atmospheric perspective, provides some sense of the energy contained in the fluctuating flow field, and is well-suited for presenting the utility of the total variation method below. Additionally, TI is a quantity frequently used in the wind energy community to characterize wind plant operating conditions and structural loading of wind turbines (Dimitrov et al., 2018; Kelly et al., 2014) and is often accessible through instrumentation on met masts or wind turbine nacelles making it an appropriate choice for the current demonstration.

Raw data used to demonstrate the current methods include high-frequency (20 Hz) observations of u and θ reported by the met mast between March and December 2009. Wind speed and direction data were binned to a temporal resolution of 1 min, from which mean and standard deviations were calculated. Turbulence intensity in each bin is estimated as the ratio of the retained 1-min statistics for wind speed as $TI = \sigma_u/u$. As with most field observations, data availability from each channel is less than 100%, as instruments require maintenance, loose connectivity to data acquisition systems, or shut down to prevent damage under certain conditions. Binning the data into 1-min periods smooths the observed time series of wind speed and direction, and reduces the noise reported by the cup anemometer and wind vane.

Additional quality-control steps for the data include omitting any 1-min period any of the data channels are not correctly reported from further consideration. Any time stamp associated with wind speeds less than 1 m/s, when wind speed observations reported by cup anemometers and wind vanes are not considered to be reliable (IEC, 2005), are also removed from the data set. Fig. 1(b) shows data availability of the record as a percent of the total number of data possible per day. The final quality-control step implemented for the current study is to exclude data that are not part of any continuous set of observations of at least 60 min. The current method searches continuous data samples to identify atmospheric conditions and events of interest. Rather than infill or interpolate data, periods with missing values are simply excluded from consideration.



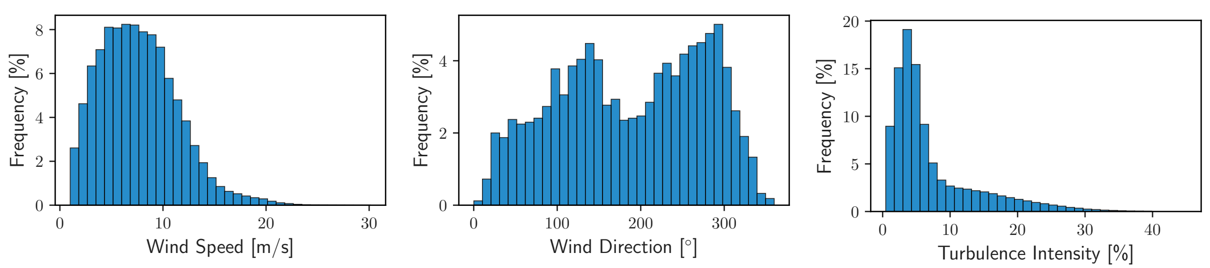
(a) Lillgrund layout, wind turbines are filled markers, met mast is open marker

(b) Daily data availability from met mast

Figure 1. Wind turbines, met mast, and data availability from Lillgrund wind plant

3 Statistical view of atmospheric conditions

Characterization of the atmospheric conditions is most often pursued through aggregate statistics, that is without explicitly considering their evolution in time. Statistical quantities (arithmetic mean values, variances, and higher-order moments) may reflect the occurrence of infrequent events, but do not convey dynamical evolution of variables or their correlation in time. Considering atmospheric variables in terms of either their marginal distributions (as in Fig. 2 or their conditional distributions (as in Fig. 3) falls short of saying anything about the dynamics embedded in those observations. Steady-state wake models are defined to represent the time-averaged flow behind a wind turbine and many uses of high-fidelity models assume that the bulk flow speed and direction do not change in time. Effective validation of numerical modeling tools for wind energy requires that observations conform to stationary atmospheric flow (Chenge and Brutsaert, 2005; Metzger et al., 2007; Vincent et al., 2010, 2011; Guala et al., 2011) or represent a dynamic event of interest. Histograms of each of the data channels are provided in Fig. 2, showing characteristic behavior for the wind speed and turbulence intensity distributions.



(a) Histogram of wind speed

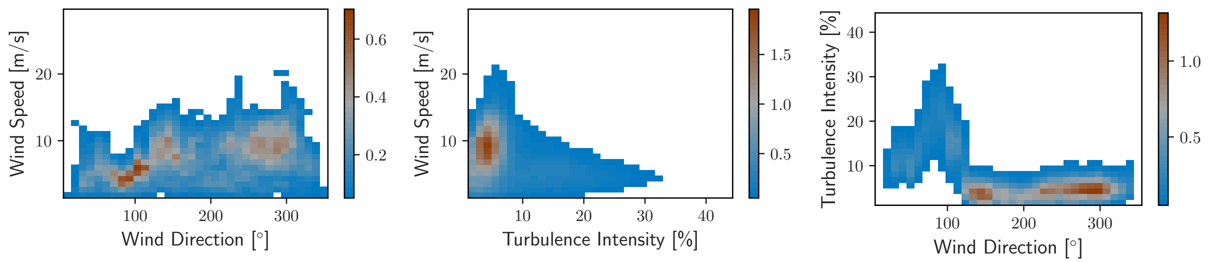
(b) Histogram of wind direction

(c) Histogram of turbulence intensity

Figure 2. Histograms of quality-controlled met mast data

The wind direction (Fig. 2(b)) shows several key features typical of atmospheric records; first, it identifies the prevailing wind directions as per the number of observations within each direction sector (10°) and, second, it shows that virtually no observations correspond with wind directly out of the north. According to IEC (2005), met masts should be placed sufficiently far from the nearest upstream obstacle, or risk introducing bias and increased uncertainty into the record. This limitation can be difficult or prohibitively expensive to accommodate due to logistical constraints, especially in offshore settings where placement is often strictly limited.

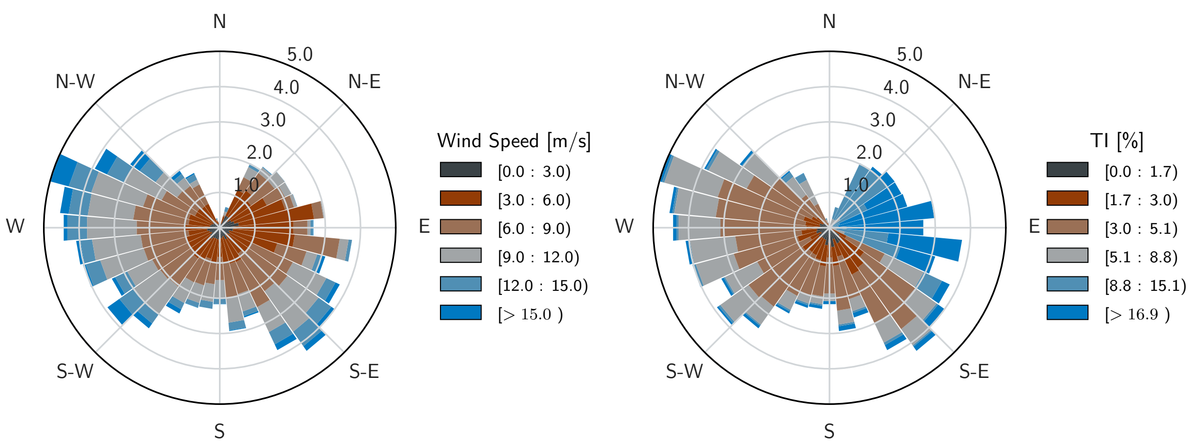
Each of the histograms in Fig. 2 categorizes a single quantity without regard to the variation of the others; each single-variable histogram effectively integrates the observations over the other two variables. More complex treatment of the data is required to take into account the simultaneous variability of more than one channel. Fig. 3 shows two-dimensional histograms with two-way permutations of the data channels. In each of the histograms, a threshold has been applied to the frequency of observations. Any bin representing less than 0.5% of the total observations has been filtered out to highlight more common conditions. Two-dimensional histograms demonstrate that the atmospheric conditions are more complex than is possible to estimate from pairwise consideration of any two of the one-dimensional histograms in Fig. 2. An observation from the two-dimensional histograms that is not immediately evident in one-dimensional histograms is that the greatest turbulence intensity



(a) Distribution of u and $\boldsymbol{\theta}$ (b) Distribution of u and TI (c) Distribution of TI and $\boldsymbol{\theta}$

Figure 3. Two-dimensional histograms of met mast data. Color information conveys percent of total observations for each pair of variable values.

comes from a single, distinct sector of wind directions. Placement of the met mast with respect to the wind turbines contributes to a sharp increase of TI in the range of 15–45% and is not typical of unobstructed measurements. Reports of high TI likely result from the introduction of turbulence to the flow by the wind turbines or wind plant from directions between 70°–110°.



(a) Wind rose (b) TI rose

Figure 4. Wind (a) and TI (b) roses from met mast data

Wind speed and TI roses contain the same information as the two-dimensional histograms from Fig. 4, but convey it on a polar projection representative of the compass, thus making them more intuitive to read for many users. Fig. 4 shows wind and TI roses for the considered data. The rose diagrams highlight directional dependence of the mapped variable. For example, Fig. 4(b) demonstrates that the greatest turbulence intensity is highly correlated with winds from the sector of 70°–110°. This is the range of directions in which the met mast is waked by the wind turbine located to the west.

4 Total variation of dynamical data

Aggregate statistical representation accounts for interdependence of the three variables considered in the current example, but cannot account for the dynamic nature of the atmosphere. A histogram, as a consequence of its composition, only denotes how frequently a given condition is observed without regard to what condition may precede or follow. The actual weather conditions could well be undergoing a dramatic change, but within any 1-min observation, the variables of interest fall within the stated bounds of a single bin within the full condition space.

An alternate path toward identifying conditions of interest for model validation or benchmarking studies comes through seeking continuous periods from the time series of observations that has properties of interest for a given study. An obvious choice would be a continuous period in which the atmospheric conditions remain statistically stationary. Statistical stationarity (i.e. time-independence of statistical quantities) is a common consideration in turbulence and atmospheric science (Chenge and Brutsaert, 2005; Metzger et al., 2007; Vincent et al., 2010, 2011; Guala et al., 2011). Stationarity is not often assumed for wind energy research and modeling applications, although it is rarely quantified or even considered in validation data. Additionally, retaining a time series allows users to leverage the interdependence of the channels within a data set by way of correlation or covariance metrics.

Quantifying the variability of a set of data must include the correlation between data channels, or risk discounting any information regarding the relationship between variables. Stated otherwise, any metric that combines the variability of each channel independently without accounting for covariance between the channels is incomplete and will not be sufficient to fully characterize the state of a given system. Therefore, a method that accounts for variation within each channel and the covariance between variables is necessary to quantify the distribution of data across multiple channels into a single metric.

Below, each data block, $\mathbf{D}$, is a selected time period and corresponds to an array of size of $[m,n]$, where m is the length of the time period — either 60 or 120 min —and n is three, corresponding to the number of variables u, θ, and TI.

$$\mathbf{D} = [u(t),\ \theta(t),\ TI(t)] \tag{1}$$

In order for the variability of each channel in D, and their respective covariances to be given equal weight, the data must be normalized to a single common range. Each variable has been normalized by its respective span and mapped to an interval determined by the range of each channel in standard deviations according to the formulation,

$$\mathbf{D}_{\text{norm}} = \frac{\mathbf{D} - \overline{\mathbf{D}}}{\sigma_{\mathbf{D}}} \tag{2}$$

In Eq. (2), the arithmetic mean and standard deviation (denoted by the overline and σ, respectively) are calculated separately for each column of $\mathbf{D}$. Normalizing data before calculating the total variation ensures that each data stream is weighted equally in the characterization of a given condition or state.

In addition to the definition of $\mathbf{D}$, a block, $\mathbf{f}$, containing objective functions of interest to apply to each of the variables in $\mathbf{D}$ is defined as,

$$\mathbf{f} = [f_u(t),\ f_\theta(t),\ f_{TI}(t)] \tag{3}$$

The difference between objective functions and their respective data will be referred to as a regularized data block, and is noted with a caret,

$$\hat{\mathbf{D}} = \mathbf{D} - \mathbf{f} \tag{4}$$

The purpose of defining an objective function block is to tune the data to show covariance specifically with respect to a desired form about which the data are regularized. Seeking stationary conditions in which minimal variation occurs in all data channels without regularization amounts to the special case of setting the function block to $\mathbf{f} = 0$ (or, more generally, when the objective function is any constant value; $\mathbf{f} = c$). The objective function block is discussed in greater detail in the following sections.

The total variation, $\mathcal{V}$, of a system is a unitless metric to quantify spread of a set of interdependent variables that accounts for autocorrelation within each channel and for covariance between channels. A covariance matrix is calculated for a subset of the data, representing a continuous period of a specified duration,

$$\mathbf{C} = \left(\frac{1}{m-1}\right)\hat{\mathbf{D}}^T\hat{\mathbf{D}} = \left(\frac{1}{m-1}\right)\begin{bmatrix} \sigma_u^2 & \sigma_u\sigma_\theta & \sigma_u\sigma_{TI} \\ \sigma_\theta\sigma_u & \sigma_\theta^2 & \sigma_\theta\sigma_{TI} \\ \sigma_{TI}\sigma_u & \sigma_{TI}\sigma_\theta & \sigma_{TI}^2 \end{bmatrix} \tag{5}$$

In Eq. (5), $\mathbf{C}$ is a square matrix of size $n \times n$ representing the covariance between any pair of data channels. The total variation, $\mathcal{V}$, of a given regularized data block, $\hat{\mathbf{D}}$, is expressed as the determinant of the respective correlation matrix,

$$\mathcal{V} = \det(\mathbf{C}) \tag{6}$$

Larger values of $\mathcal{V}$ indicate that the data points are more dispersed in the condition space. In the observational data of the atmosphere discussed here, $\mathcal{V} > 0$. The case of $\mathcal{V} = 0$ 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 multivariate data set have some history in the literature. Notable past contributions include the pooled variance method to estimate population variance from those of distinct samples Ruxton (2006), and the 'total' or 'overall' variability (Anderson, 1962; Goodman, 1968) 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 with $\mathcal{V}$, 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.

4.1 Quiescent conditions: $\mathbf{f} = c$

Fig. 5 shows the distribution of $\mathcal{V}$ dividing the data record into continuous periods of either 60 (blue) or 120 min (red). Both distributions in Fig. 5 have been limited to $\mathcal{V} \leq 0.30$ to emphasize differences between the two data block lengths. In either case, the distribution is positively skewed, and high values of $\mathcal{V}$ exist with very low frequency. Immediately visible in the histograms of $\mathcal{V}$ is that there is a range of values exhibited most commonly by the blocks of data. For data broken into 60-min

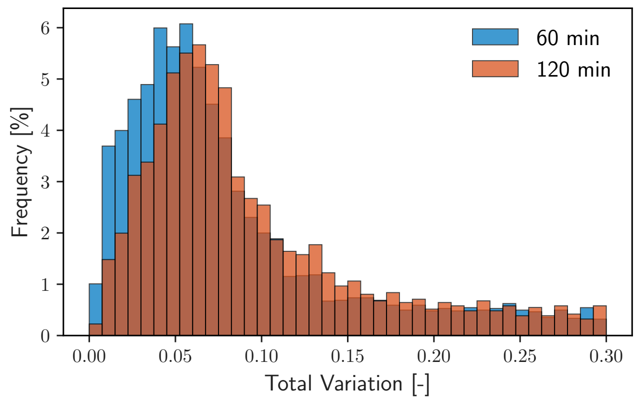
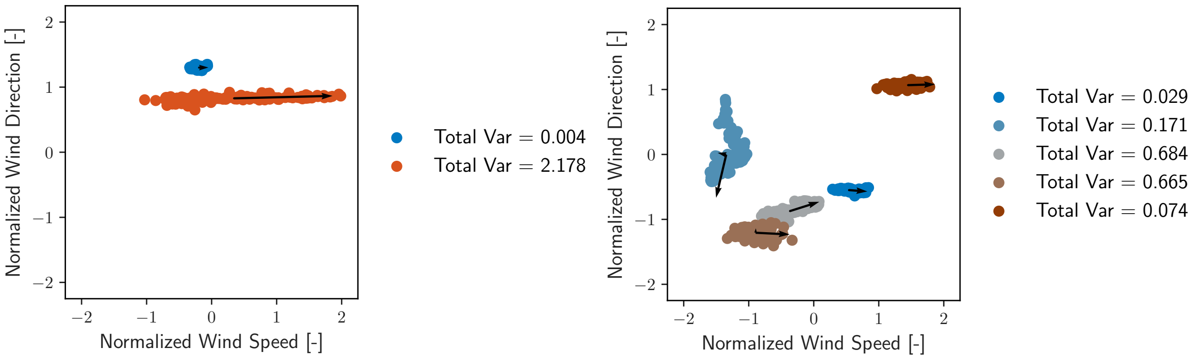


Figure 5. Distribution of $\mathcal{V}$ for data blocks of 60 or 120 min (blue and red, respectively)

periods, 35.9% of blocks have a total variation less than 0.05, whereas for data broken into 120-min periods, only 25.0% of blocks have a total variation in the same range. Although $\mathcal{V}$ is a unitless metric, its relative value does convey the degree of variation represented by all data within a respective time period. The values of $\mathcal{V}$ with the greatest frequency of occurrence is larger for periods of 120 min than for periods of 60 min. This is an expected trend because of the greater changes in atmospheric conditions that are possible within a larger window. There remains an inherent trade-off between the length of a data block and the degree of variation; longer blocks provide greater statistical convergence of $\mathbf{C}$, but risk including more dynamical variation, which contributes to higher values of $\mathcal{V}$.

Periods of time corresponding to the minimum values of $\mathcal{V}$ are those in which the total atmospheric conditions vary the least. In these periods, small values of standard deviation within each data channel as well as minimal covariance between the channels is expected. Minimal covariance between channels is equivalent to observing only stochastic, uncorrelated fluctuations in each channel. In contrast, periods corresponding to the maximum values of $\mathcal{V}$ are those in which the subset of data experiences the greatest variability, to which individual channel noise and correlated events between channels both contribute. Time periods of 120 min corresponding to the maximum (red) and minimum (blue) total variation are shown in Fig. 6(a). To provide a broader sense of how other time periods are characterized in terms of $\mathcal{V}$, five randomly selected periods of 120 min are shown in Fig. 6(b). The principal components of each data block are shown with black vectors and the total variation is listed in the legend. The figure represents each block of data as a scatter of only normalized wind speed and direction, although TI is also in the calculation of $\mathcal{V}$.

Fig. 7 shows the wind speed, direction, and turbulence intensity corresponding to the 10 periods of minimum and maximum total variation. Each variable is shown in its original (non-normalized) engineering units to provide insight into the atmospheric conditions, although they were identified using normalized data. Fig. 7(a) shows that the periods with minimal values of $\mathcal{V}$ have time series that appear constant and experience only small stochastic variations within each channel and that periods with large values of $\mathcal{V}$ exhibit more spread. For each set of time series, the extreme values are shown in the boldest color (red, blue, and



(a) Time periods exhibiting the lowest and highest values of total variation (blue and red, respectively)

(b) Scatter of observations in selected time periods

Figure 6. Scatter of data points of selected time periods within the full conditions space

gray for the wind speed, direction, and turbulence intensity, respectively) and fade to lighter colors for more moderate values of $\mathcal{V}$. Starting and ending times are not included, as Fig. 7 is intended only to demonstrate the sorting capability of the method.

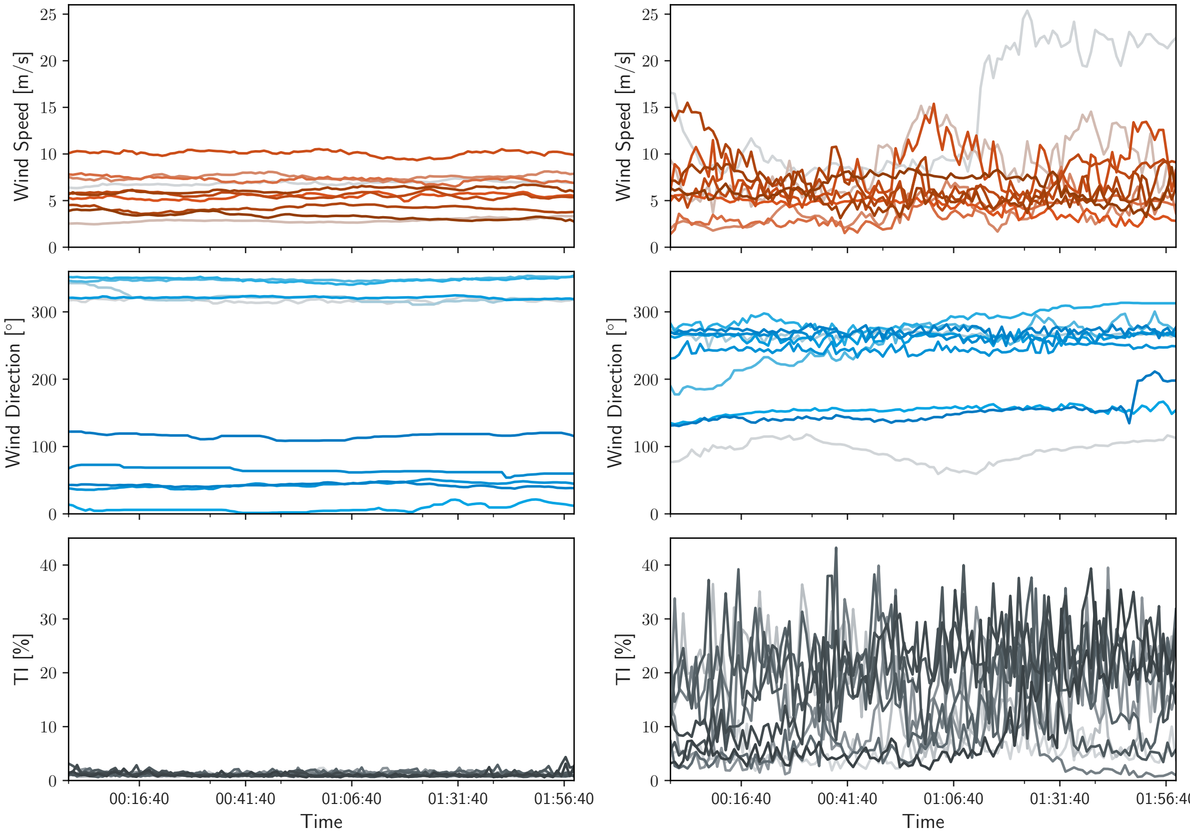
4.2 Objective conditions: $\mathbf{f} \neq 0$

Regularizing the data with respect to a set of nonzero objective functions centers the of $\mathcal{V}$ around specific conditions of interest. For example, in the case of wind plant analysis, it may be of interest to assess array performance during a wind speed ramp event or change of wind direction. Such events may be readily formulated according to accepted mathematical definitions and supplied to the total variation algorithm from Section 4. Defining specific objective functions will quantify the total variation around those conditions, which can then be used to identify the time periods that match the event of interest most closely.

An additional step is considered to sort the full data set for a more general formulation. In such a case, events of interest are defined in a suitably general formulation, and a least-squares minimization is applied to seek the relevant parameter values. In the current demonstration, function types of interest are wind speed ramps, wind speed waves, and wind direction changes, shown in the function blocks Eqs. (7), (8), and (9), respectively, distinguished with the subscripts A, B, and C.

$$f_A = \begin{cases} f_u(t) & = c_0\, t + c_1 \\ f_\theta(t) & = 0 \\ f_{TI}(t) & = 0 \end{cases} \tag{7}$$

$$f_B = \begin{cases} f_u(t) & = c_0 \sin\left(c_1\, t + c_2\right) + c_3 \\ f_\theta(t) & = 0 \\ f_{TI}(t) & = 0 \end{cases} \tag{8}$$



(a) Time periods with minimum total variation

(b) Time periods with maximum total variation

Figure 7. Time series of the 10 blocks with minimum and maximum values of $\mathcal{V}$, (a) and (b), respectively

$$f_C = \begin{cases} f_u(t) & = 0 \\ f_\theta(t) & = c_0 \arctan(c_1\, t + c_2) + c_3 \\ f_{TI}(t) & = 0 \end{cases} \tag{9}$$

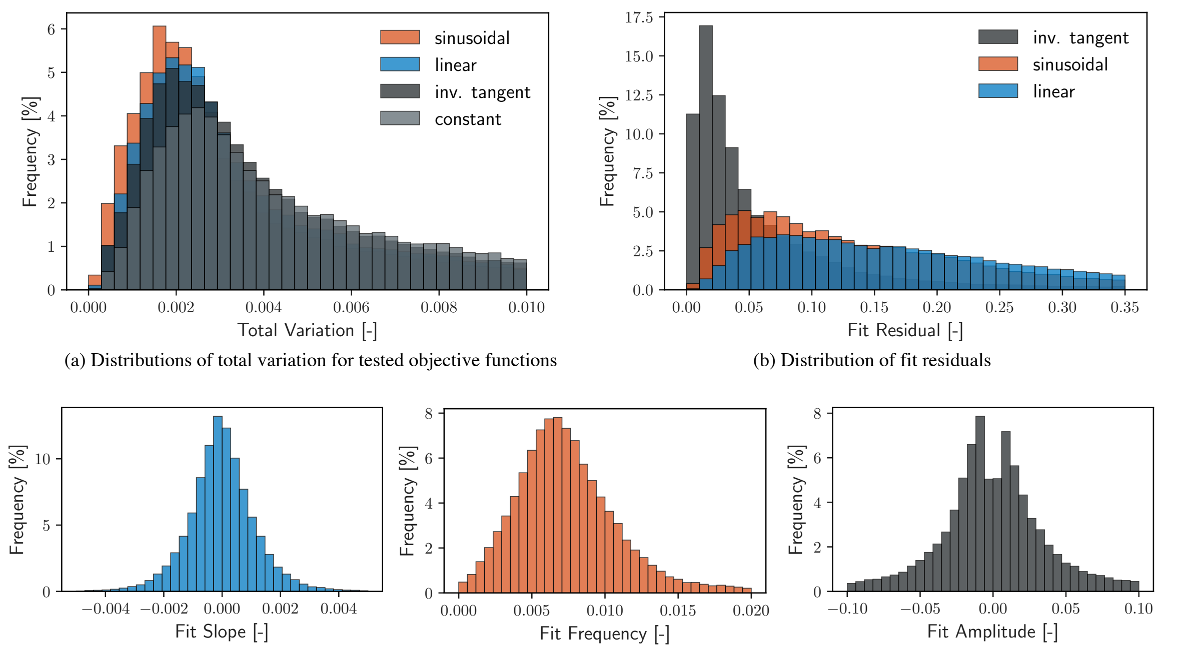
In each of the equations for f_A, f_B, or f_C, objective function parameters, c_i, are sought through least-squares minimization of the following expressions,

$$\rho = \left\| \hat{\mathbf{D}} - \mathbf{f} \right\|^2 = \begin{cases} \min \sum \left(u(t) - f_u(t, c_i)\right)^2 \\ \min \sum \left(\theta(t) - f_\theta(t, c_i)\right)^2 \\ \min \sum \left(TI(t) - f_{TI}(t, c_i)\right)^2 \end{cases} \tag{10}$$

where ρ is the least-squares fit residual. Least-squares fit parameters and the respective fit residual from each time period are retained, enabling an additional layer of filtering for conditions of interest. After objective function coefficients are determined, the total variation method is continued, yielding a value of $\mathcal{V}$ for regularized data in each time period. Regularizing the data block by subtracting away objective functions amounts to “detrending” the data such that the covariance matrix reflects correlation among the remaining data.

Fig. 8(a) compares distributions of $\mathcal{V}$ given the objective function definitions in Eq. (7), (8), and (9). The distributions indicate that the total variation can be reduced by regularizing data around generalized sinusoidal (red), linear (blue), and inverse tangent (black) functions as compared to the case where $\mathbf{f} = 0$ (gray). However, the reduction in $\mathcal{V}$ for the full data set is caused by the general definitions of the objective functions. Defining the coefficient values ahead of time would likely *increase* the average value and spread of $\mathcal{V}$; 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.

Noted earlier, the additional step of least-squares minimization provides a fit residual for each time period under consideration, shown in Fig. 8(b). Fit residuals indicate the goodness of fit of a given time period to the specified objective function forms. The distributions in Fig. 8(b) suggest that inverse tangent and sinusoidal functions fit the data with less residual error, ρ, than a linear objective function. This is likely caused by the additional objective function parameters (degrees of freedom) available for tuning the minimization.



(a) Distributions of total variation for tested objective functions

(b) Distribution of fit residuals

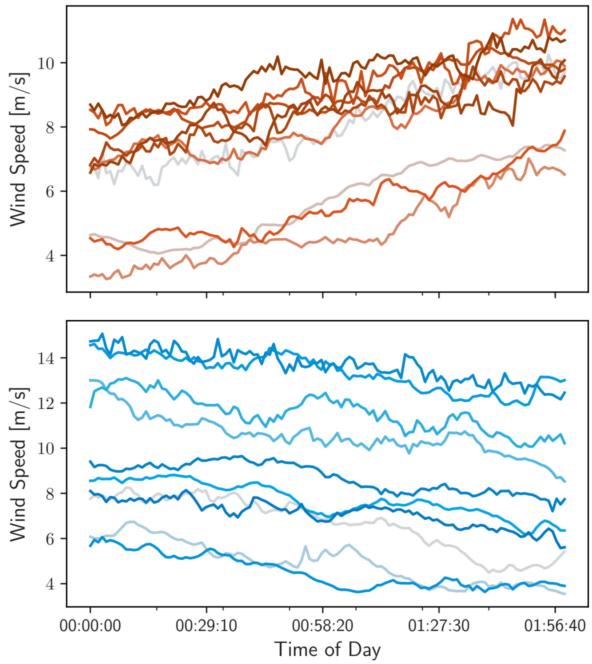
(c) Distribution of slope, c_0 from eq. (7)

(d) Distribution of frequency, c_0 from eq. (8)

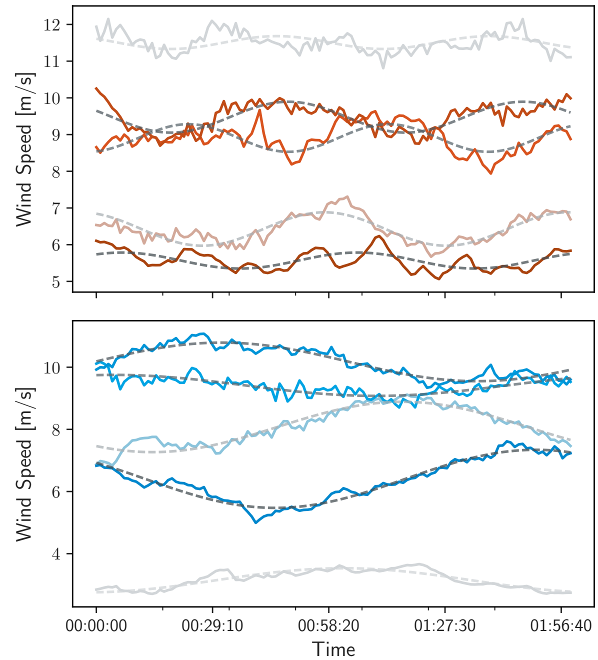
(e) Distribution of amplitude, c_0 from eq. (9)

Figure 8. Distributions of selected quantities for selected objective functions

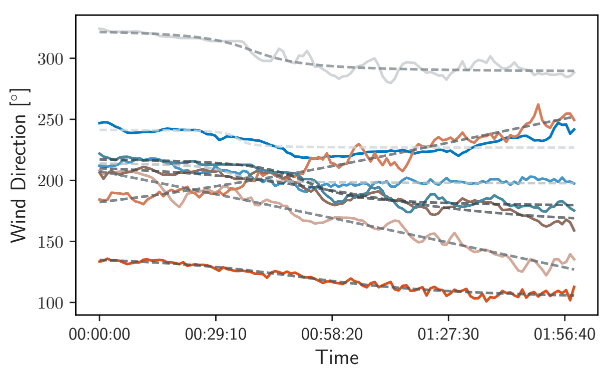
Adding an auxiliary step to the search process of least-squares minimization to a given objective function quantifies the goodness of fit of each data block and can return the parameter values necessary for the desired fit. For example, a least-squares fit to a linear relationship for any data channel will provide values of slope and offset as well as a residual value indicating the quality of the fit. In this way, the data provide alternative values for which sorting may be applied in addition to the total variation.



(a) Most linear time series from met mast



(b) Wind speed wave time series from met mast



(c) Inverse tangent time series from met mast

Figure 9. Examples of time series identified by calculating covariance matrix around linear, sinusoidal, and inverse tangent objective functions

Figs 9(a) and 9(b) show a selection of periods with minimal total variation around linear and sinusoidal objective functions of wind speed, corresponding to wind speed ramps and waves, respectively. Selection of the wind speed ramps in Fig. 9(a) are conditioned to have the minimal total variation, minimal fit residual, and maximum absolute values of slope. These are the time periods in which the wind speed ramps are simultaneously the most well-behaved (i.e. minimal fit residual) and most intense (i.e. greatest absolute value of slope). Similarly, the wind speed waves shown in Fig. 9(b) were selected by seeking the minimal total variation and then selecting time periods in which the fit frequency fell between desired limits. In Fig. 9(b), the top subfigure shows 120-minute time periods in which the fit frequency is in the range of $[0.015, 0.02]$ rad/s (in red), and the bottom subfigure shows time periods in which the fit frequency is in the range of $[0.0075, 0.008]$ rad/s (in blue). Frequency limits were selected arbitrarily, and are meant only as a demonstration of the method's independence of fit frequency. Fig. 9(c) applies an inverse tangent objective function to the wind direction channel while seeking constant conditions in wind speed and turbulence intensity, identifying the periods of wind direction change with minimal total variation. Direction changes were considered in an absolute sense, and Fig. 9(c) shows time periods with minimal $\mathcal{V}$ in which the absolute direction change $|\Delta\theta|$ falls in the range $([20°, 40°]$. Again, the particular magnitude of direction change selected here is arbitrary, and was selected only to demonstrate the fit to an inverse tangent objective function.

5 Sensitivity to outliers

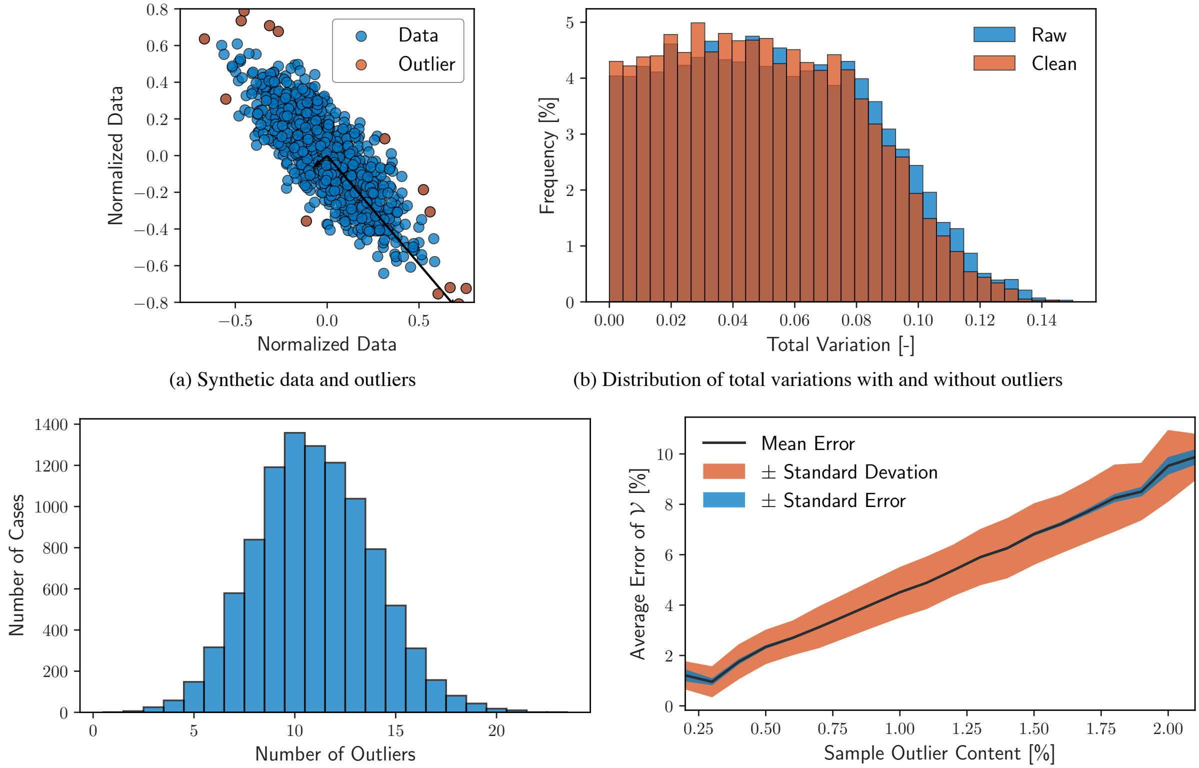
A word of caution on using the total variation to identify periods of interest: Because principal component analysis is sensitive to outliers contained in the data, the method may falsely classify a time period as having a large value of total variation due to a few spurious data points. Consideration of outliers in multivariate space requires a similar treatment as for the consideration of total variation. Seeking outlying points in each data channel individually discounts the possibility that the other data channels may be within acceptable statistical limits for the same point. Determining outliers from individual data channels further discounts any correlation that may exist between the channels. An effective means of considering outliers in multivariate data is the Mahalanobis distance, χ, which quantifies the Euclidean distance of a point from the center of a data set in terms of standard deviations (De Maesschalck et al., 2000; Hadi, 1992; Rousseeuw and Van Zomeren, 1990; Xiang et al., 2008),

$$\chi = \sqrt{(x-\mu)^T C^{-1} (x-\mu)} \tag{11}$$

The Mahalanobis distance is sought through the covariance matrix of the data, and thus accounts for interdependence of the data channels, as emphasized earlier. Setting a threshold value for the Mahalanobis distance effectively draws an n-dimensional ellipsoidal boundary around the data set in nondimensional space, outside of which data are considered invalid.

To quantify the sensitivity of $\mathcal{V}$ to the presence of outliers, 10,000 synthetic data sets are generated, and outliers are detected and eliminated. Total variation is compared for each data set before and after outlier detection/elimination. Synthetic data sets (n=2 dimensions, 1,000 points each) are normally distributed about a zero mean value with a standard deviation that is randomly assigned in the range of [0, 10]. Each data set is normalized, given a random shape parameter to stretch the data, and rotated to simulate covariance between data channels. The covariance matrix is calculated using Eq. (5) and $\mathcal{V}$ calculated as in Eq. (6). Any point with $\chi > 3$ is flagged as an outlier and eliminated. With two degrees of freedom (variables in the data block), values of $\chi > 3$ are expected to be observed with a probability of approximately 1.1% (Penny, 1996; Ben-Gal, 2005; Gellert et al., 2012). The total variation is then calculated for the cleaned data without outliers, for comparison.

Fig. 10(a) shows a single example set of synthetic data. Accepted data are shown in blue, outliers in red, and the principal components of the data are shown as the black vectors. Fig. 10(b) shows distributions of $\mathcal{V}$ before and after exclusion of



(a) Synthetic data and outliers

(b) Distribution of total variations with and without outliers

(c) Outlier detection in synthetic data

(d) Mean relative error due to the presence of outliers

Figure 10. Outlier detection and the sensitivity of $\mathcal{V}$ to outliers

outlying data identified with a threshold of χ in blue and red, respectively. As expected, the total variation of data sets without outliers is smaller than data sets before cleaning. Because of the large number of synthetic data sets considered, statistics regarding sensitivity to outliers are also within reach.

Fig. 10(c) shows the distribution of the number of detected outliers within each synthetic data set. Fig. 10(d) shows the mean relative error according to the number of detected outliers according to

$$\varepsilon = \frac{\mathcal{V}_{\text{raw}} - \mathcal{V}_{\text{clean}}}{\mathcal{V}_{\text{raw}}} \tag{12}$$

where the subscripts denote the presence and absence of outliers (raw and clean, respectively). Uncertainty of the error is shown as the shaded bands around the mean relative error. The red band indicates the standard deviation of the relative error (σ_ε) and the blue band denotes the standard error ($\sigma_\varepsilon / N_{\text{outliers}}$). The roughly linear relationship shown in Fig. 10(d) indicates that one could expect an increase in error of approximately 4% for each additional percent outlier content of a given data set.

It should be noted that the present error analysis is not expected to yield identical results for atmospheric data. Observations of wind speed, direction, and turbulence intensity can vary considerably during any given period as part of the normal development of weather patterns. Mentioned briefly in the introduction, quality control of met mast and SCADA data is an active research topic and is beyond the scope of the current method development. However, it should be clear from the sensitivity analysis undertaken here that a careful quality control process should be applied before calculation of the total variation.

6 Conclusions

The definition of high-value conditions for wind plant analysis is ultimately up to the user, but may not conform to the most frequently observed state. For example, it may be of greater concern to wind plant developers, owners, or operators to be able to validate models where wake losses are greatest or during ramps of wind speed. These conditions may be more relevant to control or curtailment actions of wind plants, and may have a greater impact on the return on investment of wind energy assets.

Identification of continuous time periods that conform to conditions of interest is not intuitive through aggregate statistics, such as measures of central tendency or even joint probability distributions. The method to quantify the total variation of a multivariate data set described earlier provides a computationally economical means of parsing large and complex data sets, and includes a mathematically robust approach to sorting with respect to a desired condition or objective function. 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. Normalizing the data makes combining disparate types of data into a single metric possible and meaningful.

The total variation method for seeking conditions of interest has applications far beyond the demonstration undertaken in the current work. Once properly classified, any number of detection and forecasting models may be trained and thoroughly validated. Collecting time periods containing similar dynamical events opens a path forward for more advanced analyses, such as modal decomposition methods and reduced order modeling. Extreme atmospheric events, as from the International Electrotechnical Commission (IEC) Standard for Wind Turbine Design (IEC, 2005), have well-defined characteristic functions and would thus fit well with the method explored in this article. After detection, wind turbine structural dynamics can be coupled to dynamical atmospheric events to produce robust and accurate control and cost models.

The total variation method explored here details identification and characterization of time series data from met masts only. Validation of high-fidelity wind plant models frequently relies on some form of operational data, most often power production or some integrated statistic of wind plant performance. SCADA signals and power production or fault events could readily be identified with the total variation method. A further extension of the method would be to add functionality that accounts for spatial variation of operational data within a wind plant. A spatial aspect to the total variation method would augment the process to be able to detect and characterize the movement of weather fronts through a wind plant or cases in which wake losses are particularly significant and heterogeneous.

Acknowledgements. This work was authored by the National Renewable Energy Laboratory, operated by Alliance for Sustainable Energy, LLC, for the U.S. Department of Energy (DOE) under Contract No. DE-AC36-08GO28308. Funding provided by the U.S. Department of Energy Office of Energy Efficiency and Renewable Energy Wind Energy Technologies Office. The views expressed in the article do not necessarily represent the views of the DOE or the U.S. Government. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a nonexclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this work, or allow others to do so, for U.S. Government purposes. Data was furnished to the authors under an agreement between the National Renewable Energy Laboratory, Siemens Gamesa Renewable Energy A/S, and Vattenfall. Data and results used herein do not reflect findings by Siemens Gamesa Renewable Energy A/S and Vattenfall. Additional thanks to the National Renewable Energy Laboratory Wake Squad for the musings and dialogue that ultimately led to this work getting started (and finished). Special thanks to Tony Martinez for countless discussions on everything from numerical methods to physical interpretations, editing, and computational assistance with volume rendering. Bridging the gap between my own turbulence experience and Mike Optis' atmospheric perspective essentially framed the discussion and motivation of the work.

References

Ali, N., Hamilton, N., Calaf, M., and Cal, R. B.: Turbulence kinetic energy budget and conditional sampling of momentum, scalar, and intermittency fluxes in thermally stratified wind farms, Journal of Turbulence, pp. 1–32, 2019.

Anderson, T. W.: An introduction to multivariate statistical analysis, Tech. rep., Wiley New York, 1962.

Barthelmie, R., Crippa, P., Wang, H., Smith, C., Krishnamurthy, R., Choukulkar, A., Calhoun, R., Valyou, D., Marzocca, P., Matthiesen, D., et al.: 3D wind and turbulence characteristics of the atmospheric boundary layer, Bulletin of the American Meteorological Society, 95, 743–756, 2014.

Barthelmie, R., Churchfield, M. J., Moriarty, P. J., Lundquist, J. K., Oxley, G., Hahn, S., and Pryor, S.: The role of atmospheric stability/turbulence on wakes at the Egmond aan Zee offshore wind farm, in: Journal of Physics: Conference Series, vol. 625, p. 012002, IOP Publishing, 2015.

Belušić, D. and Mahrt, L.: Is geometry more universal than physics in atmospheric boundary layer flow?, Journal of Geophysical Research: Atmospheres, 117, 2012.

Ben-Gal, I.: Outlier detection, in: Data mining and knowledge discovery handbook, pp. 131–146, Springer, 2005.

Bossavy, A., Girard, R., and Kariniotakis, G.: Forecasting ramps of wind power production with numerical weather prediction ensembles, Wind Energy, 16, 51–63, 2013.

Chenge, Y. and Brutsaert, W.: Flux-profile relationships for wind speed and temperature in the stable atmospheric boundary layer, Boundary-Layer Meteorology, 114, 519–538, 2005.

Cutler, N., Kay, M., Jacka, K., and Nielsen, T. S.: Detecting, categorizing and forecasting large ramps in wind farm power output using meteorological observations and WPPT, Wind Energy: An International Journal for Progress and Applications in Wind Power Conversion Technology, 10, 453–470, 2007.

De Maesschalck, R., Jouan-Rimbaud, D., and Massart, D. L.: The mahalanobis distance, Chemometrics and intelligent laboratory systems, 50, 1–18, 2000.

Dimitrov, N. K., Kelly, M. C., Vignaroli, A., and Berg, J.: From wind to loads: wind turbine site-specific load estimation with surrogate models trained on high-fidelity load databases, Wind Energy Science, 3, 767–790, 2018.

Eaton, M. L.: Multivariate statistics: a vector space approach., JOHN WILEY & SONS, INC., 605 THIRD AVE., NEW YORK, NY 10158, USA, 1983, 512, 1983.

Ferreira, C., Gama, J., Matias, L., Botterud, A., and Wang, J.: A survey on wind power ramp forecasting., Tech. rep., Argonne National Lab.(ANL), Argonne, IL (United States), 2011.

Ferreira, C., Gama, J., Miranda, V., and Botterud, A.: Probabilistic ramp detection and forecasting for wind power prediction, in: Reliability and risk evaluation of wind integrated power systems, pp. 29–44, Springer, 2013.

Fulcher, B. D.: Feature-based time-series analysis, in: Feature Engineering for Machine Learning and Data Analytics, pp. 87–116, CRC Press, 2018.

Gamage, N. and Hagelberg, C.: Detection and analysis of microfronts and associated coherent events using localized transforms, Journal of the atmospheric sciences, 50, 750–756, 1993.

Gellert, W., Hellwich, M., Kästner, H., and Küstner, H.: The VNR concise encyclopedia of mathematics, Springer Science & Business Media, 2012.

Goodman, M.: 242. Note: A Measure of'Overall Variability'in Populations, Biometrics, pp. 189–192, 1968.

Guala, M., Metzger, M., and McKeon, B. J.: Interactions within the turbulent boundary layer at high Reynolds number, Journal of Fluid Mechanics, 666, 573–604, 2011.

Hadi, A. S.: Identifying multiple outliers in multivariate data, Journal of the Royal Statistical Society: Series B (Methodological), 54, 761–771, 1992.

Hannesdóttir, A. and Kelly, M.: Detection and characterization of extreme wind speed ramps, Wind Energy Science Discussions, 2019, 1–18, 2019.

Hansen, K. S., Barthelmie, R. J., Jensen, L. E., and Sommer, A.: The impact of turbulence intensity and atmospheric stability on power deficits due to wind turbine wakes at Horns Rev wind farm, Wind Energy, 15, 183–196, 2012.

Holtslag, A. A. and Nieuwstadt, F. T.: Scaling the atmospheric boundary layer, Boundary-Layer Meteorology, 36, 201–209, 1986.

Hristov, T., Friehe, C., and Miller, S.: Wave-coherent fields in air flow over ocean waves: Identification of cooperative behavior buried in turbulence, Physical review letters, 81, 5245, 1998.

IEC, I.: 61400-1: Wind turbines part 1: Design requirements, International Electrotechnical Commission, p. 177, 2005.

Kaimal, J., Wyngaard, J., Haugen, D., Coté, O., Izumi, Y., Caughey, S., and Readings, C.: Turbulence structure in the convective boundary layer, Journal of the Atmospheric Sciences, 33, 2152–2169, 1976.

Kang, Y., Belušić, D., and Smith-Miles, K.: Detecting and classifying events in noisy time series, Journal of the Atmospheric Sciences, 71, 1090–1104, 2014.

Kang, Y., Hyndman, R. J., and Smith-Miles, K.: Visualising forecasting algorithm performance using time series instance spaces, International Journal of Forecasting, 33, 345–358, 2017.

Kelly, M., Wyngaard, J. C., and Sullivan, P. P.: Application of a subfilter-scale flux model over the ocean using OHATS field data, Journal of the Atmospheric Sciences, 66, 3217–3225, 2009.

Kelly, M., Larsen, G., Dimitrov, N. K., and Natarajan, A.: Probabilistic meteorological characterization for turbine loads, in: Journal of Physics: Conference Series, vol. 524, p. 012076, IOP Publishing, 2014.

Kumar, P. and Foufoula-Georgiou, E.: Wavelet analysis for geophysical applications, Reviews of geophysics, 35, 385–412, 1997.

Kumar, V., Kleissl, J., Meneveau, C., and Parlange, M. B.: Large-eddy simulation of a diurnal cycle of the atmospheric boundary layer: Atmospheric stability and scaling issues, Water resources research, 42, 2006.

Larsen, T. J., Madsen, H. A., Larsen, G. C., and Hansen, K. S.: Validation of the dynamic wake meander model for loads and power production in the Egmond aan Zee wind farm, Wind Energy, 16, 605–624, 2013.

Lilly, J. M.: Element analysis: a wavelet-based method for analysing time-localized events in noisy time series, Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences, 473, 20160 776, 2017.

Mathis, R., Hutchins, N., and Marusic, I.: Large-scale amplitude modulation of the small-scale structures in turbulent boundary layers, Journal of Fluid Mechanics, 628, 311–337, 2009.

Metzger, M., McKeon, B., and Holmes, H.: The near-neutral atmospheric surface layer: turbulence and non-stationarity, Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 365, 859–876, 2007.

Penny, K. I.: Appropriate critical values when testing for a single multivariate outlier by using the Mahalanobis distance, Journal of the Royal Statistical Society: Series C (Applied Statistics), 45, 73–81, 1996.

Preston, D., Protopapas, P., and Brodley, C.: Discovering arbitrary event types in time series, Statistical Analysis and Data Mining: The ASA Data Science Journal, 2, 396–411, 2009.

Rousseeuw, P. J. and Van Zomeren, B. C.: Unmasking multivariate outliers and leverage points, Journal of the American Statistical association, 85, 633–639, 1990.

Ruxton, G. D.: The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test, Behavioral Ecology, 17, 688–690, 2006.

Sengupta, A.: Generalized variance, Encyclopedia of statistical sciences, 2004.

Sevlian, R. and Rajagopal, R.: Wind power ramps: Detection and statistics, in: 2012 IEEE Power and Energy Society General Meeting, pp. 1–8, IEEE, 2012.

Shahabi, C. and Yan, D.: Real-time Pattern Isolation and Recognition Over Immersive Sensor Data Streams., in: MMM, pp. 93–113, 2003.

Sørensen, J. N. and Shen, W. Z.: Numerical modeling of wind turbine wakes, Journal of fluids engineering, 124, 393–399, 2002.

Sullivan, P. P., McWilliams, J. C., and Moeng, C.-H.: Simulation of turbulent flow over idealized water waves, Journal of Fluid Mechanics, 404, 47–85, 2000.

Sun, J., Nappo, C. J., Mahrt, L., Belušić, D., Grisogono, B., Stauffer, D. R., Pulido, M., Staquet, C., Jiang, Q., Pouquet, A., et al.: Review of wave-turbulence interactions in the stable atmospheric boundary layer, Reviews of geophysics, 53, 956–993, 2015.

Taswell, C.: Handbook of wavelet transform algorithms, 2001.

Vincent, C., Giebel, G., Pinson, P., and Madsen, H.: Resolving nonstationary spectral information in wind speed time series using the Hilbert–Huang transform, Journal of Applied Meteorology and Climatology, 49, 253–267, 2010.

Vincent, C. L., Pinson, P., and Giebela, G.: Wind fluctuations over the North Sea, International Journal of Climatology, 31, 1584–1595, 2011.

Wasserman, L.: All of statistics: a concise course in statistical inference, Springer Science & Business Media, 2013.

Wilks, S. S.: Certain generalizations in the analysis of variance, Biometrika, pp. 471–494, 1932.

Wyngaard, J. C.: Turbulence in the Atmosphere, Cambridge University Press, 2010.

Xiang, S., Nie, F., and Zhang, C.: Learning a Mahalanobis distance metric for data clustering and classification, Pattern recognition, 41, 3600–3612, 2008.

Zhang, J., Florita, A., Hodge, B.-M., and Freedman, J.: Ramp forecasting performance from improved short-term wind power forecasting, in: ASME 2014 international design engineering technical conferences and computers and information in engineering conference, pp. V02AT03A022–V02AT03A022, American Society of Mechanical Engineers, 2014.