Lossless Compression of Wind Plant Data

Henry Louie, *Member, IEEE*, Agnieszka Miguel *Member, IEEE*

***Abstract*—Substantial quantities of wind plant data are being accumulated as interest and investment in renewable energy grows. These data sets can approach tens of terabytes in size, making their management, storage, manipulation, and transmission burdensome. Lossless compression of the data sets can mitigate these challenges without sacrificing accuracy. This paper develops and analyzes lossless compression algorithms that can be applied to data used in integration studies and data used in wind plant monitoring and operation. The algorithms exploit wind speed-to-wind power relationships, and the temporal and spatial correlations in the data. The Shannon entropy of wind power and speed data is computed to gain insight on the uncertainty of wind power and speed and to benchmark performance of the compression algorithms. The algorithms are applied to the National Renewable Energy Laboratory's Western and Eastern Data Sets and to actual wind turbine data. The resulting compression ratios are up to 50 percent higher than those obtained by direct application of off-the-shelf lossless compression methods.**

***Index Terms*—Data compression, entropy, wind energy, wind power.**

I. Introduction

WIND plant data sets play an important role in several data-driven tasks, such as wind integration studies and wind plant operation [1]. In the context of this paper, a wind plant data set broadly refers to a data set containing one or more synchronized time series of wind power and wind speed for one or more locations. Other related data fields such as wind speed at various heights or turbine-specific parameters such as rotor RPM, nacelle position and turbine availability may also be present [2]–[4]. The wind resource data used in integration studies and wind turbine SCADA data are examples of wind plant data sets.

Wind plant data sets often contain time series for many locations, and can be classified according to the regularity of the spacing of the locations. If the time series correspond to locations spaced at consistent intervals, then the data set is classified as gridded or raster data. If the data set contains time series for irregularly spaced locations, then the data set is classified as point data.

Both gridded and point data sets can be voluminous in size, making their management, storage, transmittal and manipulation burdensome. For example, the National Renewable Energy Laboratory (NREL) Western Data Set with all data fields, including atmospheric quantities, is over 24 terabytes in size when stored in the standard netCDF format [3], [5], [6]. The companion Eastern Data Set is of similar size [4]. Some of the the challenges encountered in working with such large data sets are documented in [3] and [4]. Remarkably, in some cases the data were shared via physical delivery on hard drives rather than through electronic transmittal due to their large size.

A common approach to handling large data sets is to compress the data. Algorithms are used to eliminate some of the inherent redundancies in the data, thereby reducing the overall file size [7]. The data can then be decompressed as needed to retrieve the original data set. An example of data compression from the atmospheric science community is the use of the GRIB2 (Gridded Binary 2) format by the World Meteorological Organization [8]. GRIB2 compresses gridded data sets by removing spatial redundancies.

The compression of wind plant data sets is currently limited. For example, the Western Data Set is accessed uncompressed in the .csv format [9] and Eastern Data Set is compressed using only the Deflate (.zip) method [10]. Wind turbine data are often compressed within the process information software using methods such as dead-band and swinging door [11], [12]. However these methods are lossy and do not allow for the exact replication of the data. Within the field of power engineering, data compression has applications in power system monitoring—for example, compressing data from phasor measurement units [13]—power quality [14], and power system protection and stability [15].

The lossless compression of wind plant data sets poses a new challenge that is of interest to the wind industry [16]. To achieve high levels of compression, an algorithm is needed that is tailored to the inherent characteristics of wind plant data. No such algorithm has been developed for wind power or hub-height wind speed data. The task is challenging, as it will be shown that wind plant data has a high level of Shannon entropy caused by the general erratic nature of wind power and wind speed.

Compressing wind plant data sets has several benefits. Compressed data sets can be shared by the research and industry community quickly; data storage costs are decreased, along with the costs of managing the stored data; and the data may be more amenable to data mining. This could lead to the creation of easily accessible expanded wind resource data sets and the ability to exactly and completely reconstruct past wind turbine data for post-mortem analyses.

The purpose of this paper is to develop and evaluate lossless compression algorithms for point and gridded wind plant data sets. In particular, the wind power and wind speed data fields are compressed. Concepts from information theory such as Shannon entropy are applied to the data to gain insight about the characteristics of wind resource data and the limits of the performance of compression algorithms. The algorithms are applied to the NREL Eastern [10] and Western [9] Data Sets and actual wind turbine data [17] for validation and

Manuscript submitted January 2, 2012.

The authors are with the Department of Electrical and Computer Engineering, Seattle University, Seattle, WA, 98122, USA (e-mail: louieh@seattleu.edu; amiguel@seattleu.edu).

This work is funded by the Washington NASA Space Grant Consortium.

comparison with other lossless compression methods.

The remainder of this paper is arranged as follows. The data sets considered in this work are described in Section II. Section III provides background theory on lossless data compression. The Shannon entropy characteristics of the data sets are examined in Section IV. Section V and Section VI develop and evaluate lossless compression algorithms for point and gridded data sets, respectively. Conclusions and future outlook are provided in Section VII.

II. Data Set Descriptions

There are three sources of data considered in this paper: synthesized wind resource data from the NREL Eastern [10] and Western [9] Data Sets and high-frequency measured wind turbine data [17]. The compression algorithms described in this paper can be applied to other data sets and the generalization of the results are described where appropriate.

The first two data sets, hereafter referred to as the Eastern and Western Data Sets, were created by AWS Truepower and 3TIER, respectively. They were created in support of two comprehensive renewable energy integration studies on the eastern and western portions of the United States [3]–[5], [18], [19]. The large scope, use of different state-of-the-art methods and public availability make these data sets good candidates for analysis, with widely applicable results.

The Eastern and Western data sets were created from numerical weather simulations. The atmospheric data produced from these simulations were post-processed to create time series of power output by hypothetical wind plants at various locations. The data in each set correspond to wind power and wind speed over a three year period with ten-minute granularity—a total of 157 824[1] intervals. Each data set used different post-processing methods and considered wind plants of different specifications. It will be shown that the differences in capacity and precision have a profound impact on entropy and compressibility.

The Eastern Data Set contains time series for 1326 hypothetical wind plants of heterogeneous geographic footprints, capacities and wind turbine power curves. The irregular spacing of the wind plants makes the Eastern Data Set a point data set. The capacity for the wind plants ranges from 100 MW to 1435 MW, with a median capacity of 370 MW. The data are stored at a precision of 100 kW and 0.001 m/s.

The Western Data Set contains time series for approximately 32 000 hypothetical wind plants. The locations were selected from a gridded data set of over 1.2 million locations. Due to this selection process, the 32 000 locations are not all contiguous. This is functionally equivalent to a gridded data set with null or missing values for the locations without selected wind plants. Each wind plant is identical and has a 30 MW capacity. The data are stored at a precision of 1 kW and 0.01 m/s.

The third data set is measured data from a wind turbine. It contains wind power and speed values sampled at a rate of 1 Hz from a 300 kW Nordtank wind turbine over a period of eight hours. It is shown that this limited sampling window affects the entropy and compressibility of the data. The turbine is located at the Norrekaer Enge wind plant in Denmark [17]. The data are stored at a precision of 0.1 kW and 0.01 m/s.

The range and precision of the data fields in all three data sets are such that they can be parsimoniously stored in the 16 bit binary file format. This is the assumed original file format used in the remainder of this paper for compression ratio calculations. It is notable that the assumption makes for more conservative compression ratio values.

III. Lossless Data Compression

Data compression—also known as data coding—is the process of reducing the size of a digital representation of data such as text, image or audio signal [7]. In this paper, the data compressed are point and gridded wind power and wind speed time series values. Compression is achieved by reducing the redundancy in the data set, while maintaining most or all the original information.

Compression methods are classified as lossless or lossy. Lossless compression guarantees the recovery of an exact replica of the original data. Except in special cases, lossless compression provides only limited compression ratios. This paper considers lossless data compression only as this alone is a rich subject. Lossy compression will be explored in future research.

An important figure of merit in data compression is the compression ratio. The compression ratio is defined as the ratio of the file size prior to compression to the file size after compression. The design of compression algorithms involves trade-offs among such factors as the compression ratio and the time and computational resources required to encode (compress) and decode (decompress) the data. This paper judges performance primarily based on compression ratio. However, the described algorithms require similar computational time and resources as common off-the-shelf methods.

A. Compression Methods

Consider a data set $\boldsymbol{X}$ containing T datum points each denoted $x[\tau]$ sampled from a random variable $\tilde{x}$ with corresponding time interval τ. The potential values that a datum in the data set can obtain is known as a symbol. A convenient storage scheme is to represent each symbol in binary using a fixed nominal number of bits b, such as 8 or 16, depending on the range and precision of the data. The file size is then $T \times b$ bits.

If using a nominal number of bits is not necessary, a minimal fixed bit length scheme can be used. If the symbols are integers ranging from 0 to N-1, then the minimal number of bits used to represent each symbol using a fixed bit length scheme is:

$$b = \lceil \log_2(N) \rceil \tag{1}$$

where $\lceil \cdot \rceil$ is the ceiling function.

More advanced lossless compression methods achieve greater compression ratios by exploiting the statistical characteristics of the data. There is a multiplicity of techniques for lossless data compression [7]. Entropy coding methods assign variable length codes to different symbols based upon the frequency of occurrence of the symbol. Commonly encountered

[1] The three year period includes one 366-day year.

symbols are assigned shorter codes than less frequently occurring values. Huffman and arithmetic methods are examples of variable length coders [7], [20]. The common LZ78, LMZA (.7z) [21] and others are dictionary coders, in which sequences of symbols are stored in a dictionary. The code itself consists of pointers to locations in the dictionary.

The data can also be pre-processed or transformed and then compressed to increase compression ratios. For example, predictive pre-processing uses previous data to predict latter data. The difference between the predicted and actual value is stored, which can often be represented in fewer bits than the original data or may otherwise be more conducive to compression. Methods such as BZIP2 (.bz2) [22] and Deflate (.zip) [23] employ transformations and various encoding schemes successively.

Compression methods can also be classified as being context-based or not [24]. Briefly, context-based compression exploits the sometimes statistical dependence of adjacent symbols in data to increase the compression ratio. BZIP2, Deflate and LMZA can be considered context-based compression methods. It will be shown that wind plant data has a high level of statistical dependency and thus context-based compression is warranted.

B. Shannon Entropy

An important concept in information theory that affects the compressibility of data is Shannon entropy [25]. Shannon entropy—hereafter referred to as entropy—is a measure of the uncertainty associated with a random variable, such as wind power or speed. The zero order entropy E_0 of a random variable is:

$$E_0(\tilde{x}) = -\sum_{n=1}^{N} \rho_n \log_2 \left(\rho_n\right) \tag{2}$$

where $\tilde{x}$ is a discrete random variable that has N possible symbol values x_n with corresponding probability ρ_n [7]. The expected value of the number of bits of information in a single, specific realization of $\tilde{x}$ is E_0 bits. A data set of wind power or speed are specific realizations of sequences of their underlying random variables. Therefore, entropy can be interpreted as the expected number of bits of information per datum in a data set. Unless a context-based method is used, it is not possible to code the data using fewer bits per datum without the loss of information. As such, the compression ratio will never exceed $\frac{\bar{b}}{E_0}$, where $\bar{b}$ is the average number of bits per datum in the uncompressed format.

The zero order entropy places a strict limit on the compression ratio that can be obtained by lossless algorithms applied to a sequence of independent and identically distributed random variables. Should the values in a sequence not be independent, then higher order entropy is of interest.

Zero order entropy can be extended to the first order by considering the conditional probability of a symbol occurring given the previous symbol [26]. The first order entropy is:

$$E_1(\tilde{x}) = -\sum_{n=1}^{N} \rho_n \sum_{m=1}^{N} \rho_{n|m} \log_2 \left(\rho_{n|m}\right) \tag{3}$$

TABLE I
ESTIMATED ENTROPY OF WIND POWER AND SPEED DATA

	Western		Eastern		Turbine	
	Power	Speed	Power	Speed	Power	Speed
Min. Fixed	15	13	13	15	13	12
$E_0(X)$	10.69	10.71	11.17	13.65	11.00	9.91
$E_1(X)$	6.90	7.23	7.85	9.12	—	—
$E_0(\delta X)$	8.62	7.31	8.36	10.25	8.89	7.95

where $\rho_{n|m}$ is the probability of symbol m given that n is the previous symbol. If the first order entropy is lower than the zero order entropy, then there is dependency between adjacent values in the sequence of random variable realizations. Context-based compression methods exploit these statistical dependencies to achieve higher compression ratios.

IV. Entropy of Wind Plant Data Sets

The maximum compression ratio that can be obtained by any lossless data compression method is intimately related to the entropy of the data set. In this Section, the zero and first order entropy of wind power and speed data from the three data sets are benchmarked and analyzed.

A. Zero Order Entropy

To compute the zero order entropy of a random variable, its probability mass function (PMF) must be known, as indicated by (2). The PMFs of the random variables representing the wind power and speed at each location are not explicitly known. However, due to the large number of data points, the PMFs can be estimated by creating histograms for each data field for each location.

The average zero order entropy of the wind power and wind speed data were computed based on the estimated PMFs of each location. The entropy was computed for all 1326 locations in the Eastern Data Set and a continuous three-hour period for the wind turbine data set. The shorter period for the latter is justified by the high sample rate of 1 Hz. For convenience in comparison and computation, the average entropy of the Western Data Set was estimated from a randomly selected subset of 1326 wind plant locations. A random sample of 1326 out of 32 000 allows for greater than a 99 percent confidence interval with an error of estimation of the mean entropy of 0.1 bits per symbol.

The first row in Table I is the average number of bits per datum if the minimum fixed bit length scheme is used for each data field and data set, according to (1). The second row is the calculated average zero order entropy.

The entries of Table I illustrate the influence of data range, precision and distributions on entropy and compressibility. Starting with the first row, the wind power data in the Western Data Set has a larger minimum fixed bit length representation than the wind speed data; whereas for the Eastern Data Set the converse is true. This is caused by the relative precision of the data fields in each data set—1 kW and 0.01 m/s in the Western, 100 kW and 0.001 m/s in the Eastern. The values for the wind turbine data set are also commensurate with the range precision of its data, as described in Section II.

The zero order entropy values shown in the second row quantify the potential for non context-based lossless compression algorithms to compress the data. Reduction in file size between 15 and 40 percent are theoretically possible when compared to the minimal fixed bit length representation of data.

The greater compressibility of the wind power in the Western Data Set—from 15 to 10.69 bits per datum—when compared to the other data sets can be explained by examination of the underlying estimated PMFs. Fig. 1 shows the histograms of wind power data of specific, but generally characteristic, wind plants. The histogram from the Western Data Set exhibits clustering toward zero and rated power output. The high probability of these values tends to decrease the entropy, thus making the data more compressible. The Eastern and wind turbine data set histograms generally do not exhibit such pronounced clustering. The PMFs differ because most wind plants in the Eastern Data Set are larger—and hence more geographically diverse. This reduces the probability that all wind turbines will be idle or producing full power simultaneously. The influence of geographic diversity on compressibility is evident: for data sets of sufficient duration, increased geographic diversity tends to increase entropy and decrease compressibility.

The PMF for the wind turbine power data in the bottom of Fig. 1 is different from the others because only a narrow range of operation is captured. Other wind turbine data sets may have much different PMFs and entropy. However, a qualitative interpretation of the computed entropy is possible. Since the data captures operation between cut-in and rated wind speed, it is expected that the computed entropy of the wind power is higher when compared with data sets capturing operation below cut-in or between rated and cut-out wind speed where the power is constant.

The entropy of wind speed is predominantly influenced by the precision of the data. This is because the range and distributions are similar for all data sets. That is, all three data sets have wind speeds generally less than 50 m/s and tend to follow Rayleigh-like distributions. An empirical formula for estimating the zero order entropy of wind speed at hub height for precisions between 1 m/s and 0.001 m/s is:

$$E = \log_2(14.7D) \tag{4}$$

where D is the inverse of the decimal precision for the data. For example, applying (4) to wind speed data at a precision of 0.01 m/s sets D to 100 and yields an entropy of 10.52. This is close to the entropy of Western Data Set at 10.69 and to 10.35, which is the entropy of the of the Eastern Data Set after it has been rounded to a 0.01 m/s level of precision.

In general, the potential for non context-based compression of wind speed values is less than that for wind power. In order to achieve greater compression ratios the wind speed data can be pre-processed or a context-based compression method should be used, in which case higher order entropy becomes of interest.

B. First Order Entropy

The first order entropy of wind power and speed data is expected to be lower than the zero order entropy due to the statistical dependency between values separated by small time intervals. Computing the first order entropy requires conditional PMFs, as indicated in (3). In the data sets considered there are too few data points for a given location to meaningfully estimate conditional PMFs.

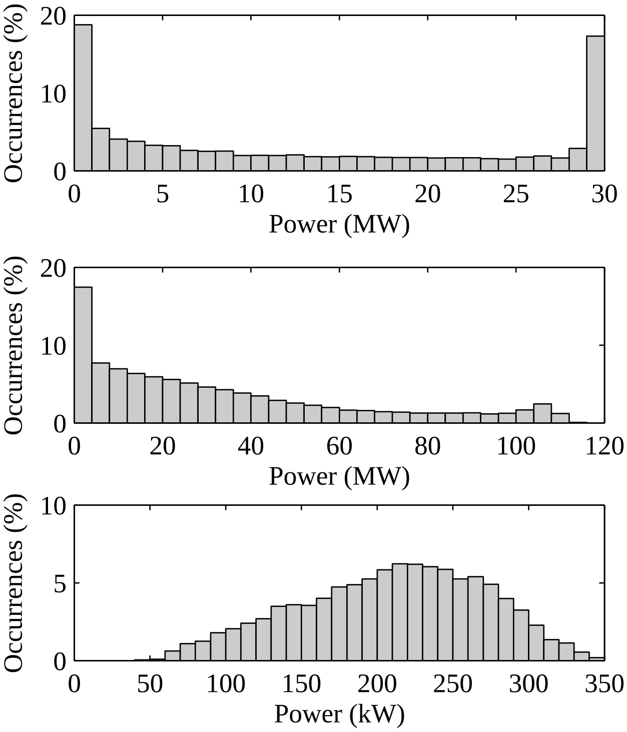

Fig. 1. Example histograms of individual wind plant locations for wind power from the Western (top) and Eastern (middle) Data Sets and for a single wind turbine (bottom).

The approach taken here for the Western and Eastern Data Sets is to concatenate the time series of 100 randomly-selected locations of similar capacities and compute the first order entropy from this expanded data set. Since sequences from different random variables corresponding to different locations are combined, the result should be interpreted as a coarse estimation only. Unfortunately, there is not a wide enough range of operation in captured in the wind turbine data set to use this approach and so it must be excluded from this analysis.

The estimated first order entropies are shown in the third row of Table I. They are lower than the zero order entropy by approximately 30 percent across all data fields. This suggests that context-based compression methods will achieve greater compression ratios than non context-based methods.

A simple method of exploiting the lower first order entropy in the data is to use a pre-processing algorithm. Conditional PMFs of wind power and speed have shown that if $x[\tau]$ has a value of x_i, then the most probable value for $x[\tau+1]$ is also x_i [27]. Therefore, first order temporal differencing can be used:

$$\delta x[\tau] = x[\tau] - x[\tau-1] \quad \forall \tau : 1 < \tau \leq T \tag{5}$$

where δ is the temporal differencing operator. Temporal differencing tends to transform the PMF of the data from those shown in Fig. 1 to a near Laplace distribution [28], which alters the entropy. The computed zero order entropy of the time-differenced data sets are provided in the last row of Table I. The Table shows that indeed the entropy has been significantly reduced when the data are pre-processed by temporal differencing. They are near, but above, the first order entropy values. The average reduction is over 35 percent when compared with the minimal fixed bit length representation, illustrating the potential for data pre-processing to increase compression ratios.

V. Lossless Compression of Point Data Sets

The insights gained from the entropy analysis in the previous Section are used to develop algorithms to compress point data sets. In particular, it was shown that pre-processing the data can effectively lower the entropy and that context-based compression methods should increase compression ratios.

Data from a single wind plant or wind turbine, or multiple wind plants or turbines at irregularly spaced intervals are point data. A simple way to compress the data is to directly use an off-the-shelf compression method, such as BZIP2. However, this approach does not fully exploit the potential of pre-processing. In this Section, a pre-processing algorithm is developed and applied to the data sets. A comparative analysis is performed against the use of off-the-shelf methods on unprocessed data. The Section concludes by analyzing the sensitivity of compression ratio to sampling frequency.

A. Algorithm Description

The proposed algorithm pre-processes wind power and speed values to reduce their entropy and allow off-the-shelf context-based methods to achieve higher compression ratios. Temporal difference coding as in (5) is used to pre-process the wind speed data. This was shown to be effective in lowering wind speed entropy in Section IV and so it is expected to enable a high compression ratio. The wind power values are pre-processed from contemporaneous wind speed values using an approximated wind plant power curve. In other words, wind speed values are used to predict wind power values. The differences between predicted and actual values are compressed. The pre-processed wind speed and power data, along with overhead information, are then compressed using an off-the-shelf method such as BZIP2. A more explicit description of the algorithm follows.

A flowchart of the algorithm is displayed in Fig. 2. Let the wind speed and wind power data for time interval τ be denoted as $v[\tau]$, $p[\tau]$, respectively. The wind speed for the first time interval $v[1]$ is stored. Temporal difference coding as in (5) is performed on the remaining wind speed data. These values, denoted as $\delta\boldsymbol{V}$, are stored.

To compress wind power data, a lookup table is used based upon an approximated wind plant power curve. The lookup table is created using a comparatively small amount of information contained as metacontent in the file to be compressed. The metacontent contains the number of wind turbines at the location and power curve of the wind turbines. It can be reasonably assumed that the wind turbine power curve is available and conforms to IEC standard 61400-12-1 [29]. As such, the power curve contains at a minimum a single power value at 0.5 m/s wind speed intervals.

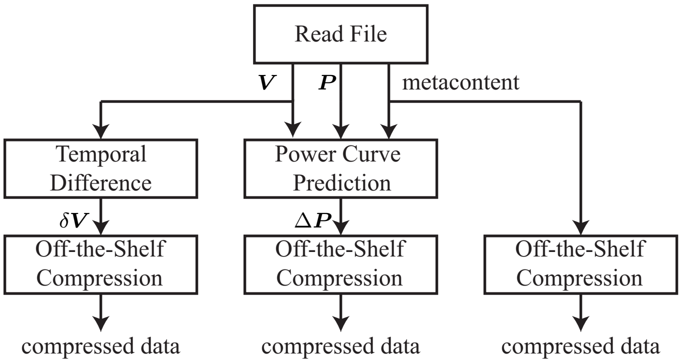

Fig. 2. Flowchart of proposed algorithm for compression of point data sets.

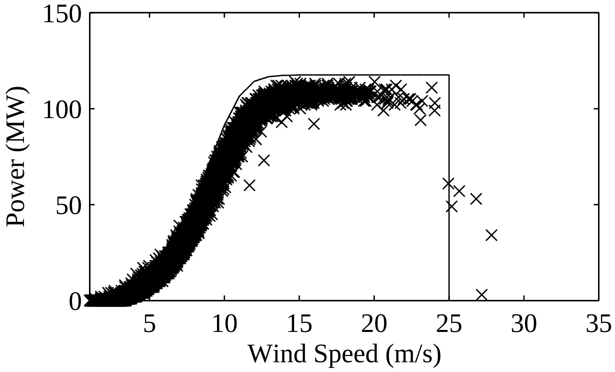

Fig. 3. Wind plant power curve and selected data points from a location in the Eastern Data Set.

The lookup table is created from the power curve in the metacontent as follows. First, the range of the power curve is extended to cover wind speeds from 0 m/s and 80 m/s. This is done by assuming zero power production for wind speeds outside the range covered by the power curve in the metacontent. The upper limit is somewhat arbitrary. It should be a value well above the highest anticipated wind speed. Next, the power curve values are linearly interpolated at intervals commensurate with the precision of the wind speed data, for example, every 0.01 m/s. If necessary, the power curve values are rounded to obtain the same precision as the wind power data stored in the data set. The resulting power curve values are then scaled by the number of wind turbines in the wind plant, as contained in the metacontent. The scaled values are stored as lookup table, arranged in order of increasing wind speed. Fig. 3 shows a graphical example of the approximated power curve for a location in the Eastern Data Set.

It is not expected that the created lookup table will result in zero prediction errors, as the wind speed-to-wind power relationship is generally not deterministic. However, as shown by the data points in Fig. 3, the errors will on average be small, and more conducive to high compression ratios. For example, the average error using this scheme on the Western and Eastern Data Sets is on the order of one percent of the capacity.

The predicted value of power for the first time interval $p^*[1]$ is then determined by evaluating the lookup table at $v[1]$. The difference between $p[1]$ and $p^*[1]$, denoted as $\Delta p[1]$, is stored. Next, the lookup table is evaluated at $v[2]$ to find the next predicted power $p^*[2]$. $\Delta p[2]$ is computed and stored. The process repeats until all power values have been considered.

The wind speed and wind power data have been transformed through the pre-processing algorithm to have lower entropy. Files containing the preprocessed wind speed data $\delta \boldsymbol{V}$ and wind power data $\Delta \boldsymbol{P}$ with the metacontent are compressed separately using an off-the-shelf lossless compression method.

The decoding of wind speed from the compressed file can be done independently from the wind power decoding, since only temporal differencing is used. This is advantageous if data mining for wind speed characteristics is needed. Decoding the wind power requires a decoded wind speed file, the re-creation of the lookup table, and adding the stored errors to the power curve-predicted values.

In summary, the algorithm requires only a small amount of overhead information to be stored—on the order of 200 bytes—is simple in implementation, only requiring linear interpolation and differencing, and can be used with a number of off-the-shelf lossless compression methods.

B. Algorithm Performance

The proposed algorithm is next applied to the data sets and evaluated through comparative analysis. The considered off-the-shelf compression methods are BZIP2, Deflate and LMZA. The methods are applied to unprocessed data and data pre-processed by the proposed algorithm. BZIP2, Deflate and LMZA are context-based methods, so the first order entropy will be exploited. The same subset of 1326 randomly-sampled locations of the Western Data Set considered in Section IV are also considered in this Section, along with the complete Eastern and wind turbine data sets.

Table II and Table III summarize the results. The shown compression ratios are with respect to an assumed nominal uncompressed 16-bit binary file format. For reference, the first row in each Table is the compression ratio if a hypothetical perfect zero order entropy encoder is used based on the average entropy values in Table I. The column headers of each data set correspond to unprocessed data ($\boldsymbol{P}$, $\boldsymbol{V}$), temporally-differenced data ($\delta \boldsymbol{P}$, $\delta \boldsymbol{V}$), and power-curve predicted data ($\Delta \boldsymbol{P}$). Temporally-differenced wind power data ($\delta \boldsymbol{P}$), while not performed in the developed algorithm, is included in Table II for comparison purposes.

Scanning down each column compares the performance of the off-the-shelf methods. Starting with Table II, it is seen that BZIP2 outperforms the other off-the-shelf methods, regardless of how or if the data were pre-processed. It averages 25 percent higher than Deflate, and 8 percent higher than LMZA.

Next, the success of the proposed pre-processing algorithm is examined. Considering the application of BZIP2 to the data, it is shown that for the Western and Eastern Data Sets, pre-processing using power curve prediction results in the highest compression ratios. From Table II, the compression ratios average 20 and 12 percent greater than the unprocessed and temporally-differenced wind power data, respectively. When compared with the typical direct application of Deflate to the data, using BZIP2 on the pre-processed data resulted in average compression increases of over 50 percent. It will be shown in Section V-C that the lower performance of power curve prediction pre-processing on the wind turbine data is due to its high frequency of sampling.

TABLE II
WIND POWER DATA COMPRESSION RATIOS

	Western			Eastern			Turbine		
Method	$\boldsymbol{P}$	$\delta \boldsymbol{P}$	$\Delta \boldsymbol{P}$	$\boldsymbol{P}$	$\delta \boldsymbol{P}$	$\Delta \boldsymbol{P}$	$\boldsymbol{P}$	$\delta \boldsymbol{P}$	$\Delta \boldsymbol{P}$
E_0	1.50	1.86	—	1.43	1.91	—	1.45	1.80	—
BZIP2	1.69	1.88	2.21	1.82	1.87	1.98	1.61	1.75	1.44
Deflate	1.39	1.54	1.70	1.35	1.53	1.49	1.29	1.45	1.23
LMZA	1.50	1.71	2.08	1.63	1.79	1.83	1.54	1.67	1.35

TABLE III
WIND SPEED DATA COMPRESSION RATIOS

	Western		Eastern		Turbine	
Method	$\boldsymbol{V}$	$\delta \boldsymbol{V}$	$\boldsymbol{V}$	$\delta \boldsymbol{V}$	$\boldsymbol{V}$	$\delta \boldsymbol{V}$
E_0	1.49	1.85	1.17	1.56	1.61	2.01
BZIP2	1.94	2.17	1.34	1.53	1.74	1.83
Deflate	1.36	1.71	1.10	1.29	1.40	1.49
LMZA	1.77	2.03	1.19	1.46	1.68	1.69

Now examining Table III for wind speed data, it is observed that again, BZIP2 achieved the highest compression ratios, whereas Deflate was the lowest, regardless of whether or not the data are pre-processed. As expected from the pervious analysis of entropy, temporal differencing of wind speed makes the data more amenable to lossless compression than the unprocessed data. When BZIP2 is used, the average increase in compression ratio due to pre-processing is 10 percent. When compared with the typical direct application of Deflate to the data, using BZIP2 on the pre-processed data resulted in average compression ratio increases of 48 percent.

C. Sensitivity to Sampling Frequency

It was shown that while power curve prediction pre-processing as described in the proposed algorithm allows for higher compression ratios for the Eastern and Western Data Sets, it was less effective for the wind turbine data. This is due to the high sample frequency of the wind turbine data.

Wind turbine data is often sampled at one second intervals and may be averaged over longer periods such as ten minutes and then stored [29]. The length of this averaging period influences the efficacy of the pre-processing algorithm, and is examined hereafter. The compression ratios resulting from the application of BZIP2 wind turbine power data for different averaging periods are shown in Fig. 4. The data correspond to a fixed eight-hour period. The decreasing trend as averaging period increases is partially an artifact of the decreasing number of data points compressed. However, it is not the absolute value of compression ratios that is of interest here; rather, it is the relative performance of the pre-processing algorithms at each averaging period that is important.

When no averaging occurs—i.e. every sample is stored—temporal differencing results in the highest compression ratio, and power curve prediction results in the lowest. However,

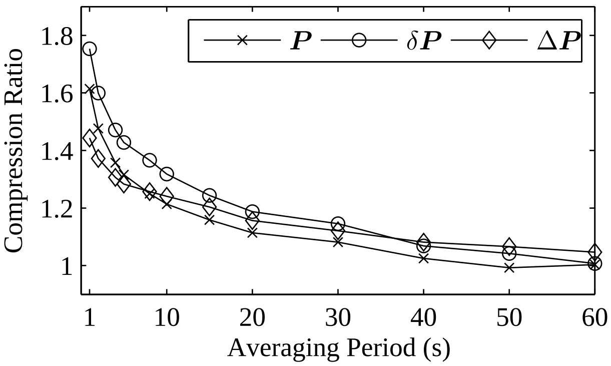

Fig. 4. Effect of sampling frequency on point data compression ratio.

as the samples are averaged over longer periods, power curve prediction pre-processing becomes more effective. When the data are stored as ten-second averages, power curve prediction pre-processing results in higher compression ratios than with unprocessed data. At 40-second averages, it overtakes the compression ratio of the temporally-differenced data. This is because temporal difference coding is most effective with no or minimal averaging, as the wind fluctuations are slow at this time scale. In addition, the power curve prediction becomes more accurate at longer time scales in part due to the low-pass filtering of the wind energy conversion system. The conclusion is that at longer averaging periods, power curve prediction results in superior compression ratios. A similar trend is observed if the data is sampled at increasing periods without averaging.

In total, the results of this Section provide several insights for the lossless compression of point wind plant data. First, if an off-the-shelf compression method is to be used on the data, it should be BZIP2 (.bz2), over Deflate (.zip) and LMZA (.7z). Second, simple temporal difference coding can be used to increase compression ratios of wind speed data and high frequency-sampled wind power data. If wind power data is sampled or averaged at rates near once per minute or slower, then power curve prediction should be used. Finally, the proposed algorithm can result in compression ratios that are up to 50 percent greater when compared to direct application of off-the-shelf methods.

VI. Lossless Compression of Gridded Data Sets

The lossless compression of gridded data sets is examined in this Section. Due to the regularity of the geographic spacing of the data, spatial redundancies can be exploited in the compression algorithm. In this Section, two existing gridded data set lossless compression algorithms are applied to wind power and wind speed data and compared to other lossless algorithms, with and without pre-processing. The goal is to evaluate gridded data set compression performance, and suggest guidelines for its implementation.

A. Algorithm Descriptions

The so-called Basic and Complex compression algorithms used in GRIB2 are commonly used on gridded meteorological data. Here they are applied and evaluated on wind power and hub height wind speed data. For convenience, a general description of the algorithms follows. A detailed description is found in [8].

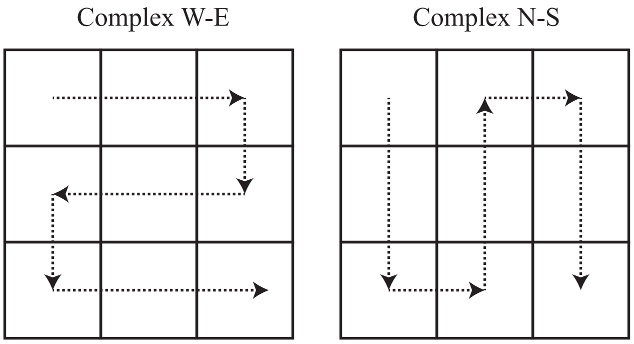

Fig. 5. Example of Complex algorithm boustrophendonical differencing in different primary directions.

Each compression algorithm begins with a user-defined set of data. This can be the entire data set or a subset thereof. It will be shown that the selection of a subset encompassing a proper geographic size is a key factor in the performance of the algorithm. The algorithms exploit spatial redundancy in the data at each time interval: the Basic algorithm subtracts the minimum value of the subset from the rest of the subset values; the Complex algorithm subtracts values at adjacent locations from each other in a boustrophendonical fashion, as shown in Fig 5, and then offsets the residuals by their minimum. The primary direction of the differencing can be West–East, or North–South. The purpose of the off-setting is to make all the values non-negative.

In both algorithms, the residuals for a given time interval are stored using a minimum fixed bit length representation as in (1). The number of bits used to represent the data at each interval is stored as overhead information. For the Complex algorithm the offset value must also be stored.

The spatial differencing used in each scheme can itself be a pre-processing algorithm if the minimum fixed bit length encoding step is eliminated, as well as offsetting the values in the Complex algorithm. This is not currently done in GRIB2, but is proposed in this paper as a viable pre-processing scheme, followed by the application of an off-the-shelf compression method to the residuals.

B. Sensitivity to Subset Size

The compression ratios yielded by the Basic and Complex algorithms are sensitive to the number of locations in the user-defined data subset. Small subsets risk lower compression ratios due to the relatively large overhead. At large subset sizes, there is less correlation of global minimums when the Basic algorithm is used, and more of a possibility for a large spatial-difference residual outlier when the Complex algorithm is used.

To determine a guideline for subset size selection, the Basic and Complex algorithms were applied to two geographically

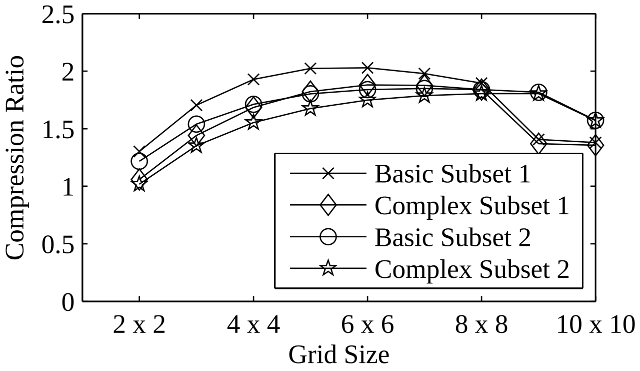

Fig. 6. Effect of subset size on compression ratio.

distinct subsets of data from the gridded Western Data Set. Information on the location of the subsets is provided in the Appendix. The subsets were increased in size ranging from 1 × 1 to 10 × 10 grids and the compression ratios calculated. The compression ratios are provided in Fig. 6. The Complex North–South algorithm is not shown for clarity, but it has similar results as the Complex West–East algorithm.

For most subsets and algorithms, there is a sharp increase in compression ratio as the subset size increases to a 6 × 6 square. For Subset 1, there is a pronounced decrease thereafter, dropping significantly at the 9 × 9 size. The algorithms applied to Subset 2 exhibit a marginal increase in compression ratio past the 6 × 6 grid sizes, and decrease sharply at the 10 × 10 size. These non-linear relationships are expected due to the aforementioned trade-offs between overhead storage requirements and spatial redundancy. The results indicate that Basic and Complex algorithms achieve the highest compression ratios when the subset dimensions are between 12 km × 12 km and 16 km × 16 km if 2 km spatial granularity is used.

C. Algorithm Performance

The Basic and Complex algorithms were separately applied to ten geographically distinct subsets of the Western Data Set. Each subset contains 36 locations arranged in 6 × 6 squares. Additional information about these subsets are found in the Appendix. The results are summarized in Table IV. The first three rows correspond to the application of the Basic and Complex algorithms as in GRIB2. The fourth through sixth rows correspond to the Basic and Complex algorithms being applied without minimal fixed bit length encoding—in other words, they pre-process the data—and then BZIP2 is applied. For comparison purposes, the final three rows correspond to the compression ratio if BZIP2 is applied to unprocessed, temporally-differenced, and power curve predicted data, as was done on the point data sets.

TABLE IV
GRIDDED DATA SET COMPRESSION RATIOS

Algorithm	Power	Speed
Basic	2.11	1.82
Complex W–E	1.94	1.94
Complex N–S	1.93	1.93
BZIP2 (Basic)	1.63	1.92
BZIP2 (Complex W–E)	1.99	2.64
BZIP2 (Complex N–S)	1.75	2.42
BZIP2 (X)	1.61	1.83
BZIP2 (δX)	1.86	2.19
BZIP2 (ΔX)	2.15	—

In regards to wind power data, Table IV shows that the Basic algorithm and the use of BZIP2 on power curve predicted data achieve the highest compression ratios at 2.11 and 2.15, respectively. They are approximately 30 percent greater than if BZIP2 was directly applied to the unprocessed data. This indicates that the existing Basic algorithm in GRIB2 can be applied to wind power data and still achieve a high compression ratio. Interestingly, the additional step of using BZIP2 after the spatial differencing in the Basic and Complex algorithms (rows four through six) generally *decreased* the compression ratios. Overall, the results show that exploiting spatial redundancy in gridded data sets leads to higher compression ratios of wind power data than if no pre-processing (row seven) or temporal differencing (row eight) is used. This supports the general claim that there is more spatial redundancy than temporal redundancy in wind power data.

For wind speed data, the use of BZIP2 on data pre-processed by the Complex West—East and Complex North—South algorithms achieved the highest compression ratios of 2.64 and 2.42, respectively. There is nearly a 45 percent increase when compared to the direct application of BZIP2 to the unprocessed data. Unlike wind power data, the use of BZIP2 on data pre-processed by the Basic and Complex algorithms *increased* the compression ratio, by 5 to 35 percent. Like wind power data, removing spatial redundancy in wind speed data results in greater compression than removing temporal redundancy.

The results for wind power and wind speed show that the primary direction of boustrophendonical spatial differencing influences the final compression ratio. That is, West–East spatial differencing outperformed North–South differencing, especially when used to pre-processing the data followed by BZIP2. This interesting result is likely due to the general west-to-east migration of weather systems in North America.

In summary, exploiting spatial redundancy in gridded data sets is beneficial, especially for wind speed data. It is more beneficial than removing temporal redundancy through differencing. For wind power data, the power curve prediction pre-processing followed by the application of BZIP2 is nearly as productive as applying the Basic Algorithm. The Basic algorithm used by GRIB2 can be applied to wind power data to obtain nearly the highest compression ratio. However, for wind speed data, the further application of BZIP2 on data pre-processed by the Complex algorithm can significantly increase compression ratios.

VII. CONCLUSION AND FUTURE OUTLOOK

Due to their large size, wind plant data sets can be cumbersome to manage, store, manipulate and transmit. This paper examined the lossless compression of wind power and wind speed data in both point and gridded data sets. The NREL Eastern and Western Data Sets and wind turbine data were used as test data sets. The zero and first order Shannon Entropy of wind power data sets were computed, providing benchmarks and insight into the development of lossless compression algorithms.

The paper proposed pre-processing algorithms that can be applied to gridded and point data sets corresponding to wind plants of varying capacities, constituent wind turbines and with different levels of precision. Sensitivity of the algorithms to sampling frequency and grid size was examined, and guidelines for their applications provided. The pre-processing algorithms resulted in improvements in compression ratios when compared with off-the-shelf compression of unprocessed data. When referenced to a parsimonious 16-bit binary file format, typical improvement is 30 percent, but can be over 50 percent in certain cases. Future directions of this work include examining the lossy compression of wind plant data.

Appendix

The gridded data used in Section VI can be accessed from [9]. The station IDs of the most northwest locations of the two subsets in Section VI-B are 16217, 2583. The station IDs of the most northwest location of the subsets used in Section VI-C are: 16217, 2583, 1642, 14606, 5003, 26714, 29041, 7352, 2024, 5385.

Acknowledgment

The authors thank Dr. Bart Niijson, Dr. Jim McCaa and Sam Cheng of 3TIER, Inc. and Shannon Hitchen of Seattle University.

References

[1] H. H. B. Ummels, J. O. Tande, A. Estanqueiro, E. Gomez, L. Söder, G. Strbac, A. Shakoor, J. C. Smith, M. Milligan, and E. Ela, "Design and operation of power systems with large amounts of wind power," Final Report of IEA Task 25, VTT Research Notes 2493, May 2009. [Online]. Available: http://www.ieawind.org/AnnexXXV.html

[2] Y. Wan, E. Ela, and K. Orwig, "Development of an equivalent wind plant power-curve," NREL, Golden, CO, Tech. Rep. NREL/CP-550-48146, May 2010.

[3] C. W. Potter, D. Lew, J. McCaa, S. Cheng, S. Eichelberger, and E. Grimit, "Creating the dataset for the Western wind and Solar Integration Study (U.S.A)," in *7th International Workshop on Large Scale Integration of Wind Power and on Transmission Networks for Offshore Wind Farms*, Madrid, Spain, May 2008.

[4] M. Brower, "Development of eastern regional wind resource and wind plant output datasets," NREL, Golden, CO, Tech. Rep. NREL/SR-550-46764, Dec. 2009.

[5] 3TIER, "Development of regional wind resource and wind plant output datasets," National Renewable Energy Laboratory, Golden, CO, Tech. Rep. NREL/SR-550-46764, Dec. 2010.

[6] Unidata, "Netcdf (network common data form)," May 2011. [Online]. Available: http://www.unidata.ucar.edu/software/netcdf/

[7] D. Salomon and G. Motta, *Handbook of Data Compression*, 5th ed. London, England, UK: Springer, 2010.

[8] C. Dey, "Guide to the WMO table driven code form used for the representation and exchange of regularly spaced data in binary form: FM 92 GRIB edition 2 layer," Digital Systems Research Center, Geneva, Switzerland, Tech. Rep., Jan. 2003.

[9] NREL. (2009, Aug.) Wind integration datasets. [Online]. Available: http://www.nrel.gov/wind/integrationdatasets/western/data.html

[10] ——. (2010, Aug.) Wind integration datasets. [Online]. Available: http://www.nrel.gov/wind/integrationdatasets/eastern/data.html

[11] S. Imtiaz, M. Shoukat Choudhry, and S. Shah, "Building multivariate models from compressed data," *Industrial Engineering Chemistry Research & Development*, vol. 46, no. 2, pp. 481–491, 2007.

[12] E. Bristol, "Data compression for display and storage," U.S. Patent 4 669 097, Oct. 1985.

[13] R. Klump, P. Agarwal, J. Tate, and H. Khurana, "Lossless compression of synchronized phasor measurements," in *Proc. PES General Meeting*, Minneapolis, MN, Jul. 2010.

[14] W. Ibrahim and M. Marcos, "Novel data compression technique for power waveforms using adaptive fuzzy logic," *IEEE Trans. Power Delivery*, vol. 20, no. 3, pp. 2136–2143, Jul. 2005.

[15] K. Mehta and B. Russell, "Data compression for digital data from power system distrubances: requirements and technique evaluation," *IEEE Trans. Power Delivery*, vol. 4, no. 3, pp. 1683–1688, Jul. 1989.

[16] B. Niijson, private communication, Jan 2010.

[17] K. Hansen and G. Larsen, "Database of wind characteristics," Oct 2011. [Online]. Available: http://www.winddata.com/

[18] GE Energy, "Western wind and solar integration study," NREL, Golden, CO, Tech. Rep. NREL/SR-550-47434, May 2010. [Online]. Available: http://www.nrel.gov/wind/systemsintegration/pdfs/2010/wwsis_final_report.pdf

[19] EnerNex Corporation, "Eastern wind integration and transmission study," NREL, Golden, CO, Tech. Rep. NREL/SR-5500-47078, Feb. 2011. [Online]. Available: http://www.nrel.gov/wind/systemsintegration/pdfs/2010/ewits_final_report.pdf

[20] D. Huffman, "A method for the construction of minimum redundancy codes," *Proc. of IRE*, vol. 40, no. 9, pp. 1098–1101, Sep. 1952.

[21] I. Pavlov, "7z format," Sept 2011. [Online]. Available: http://www.7-zip.org/7z.html

[22] J. Seward, "The bzip2 and libbzip2 official home page," Sept 2011. [Online]. Available: http://bzip.org/

[23] P. Deutsch, "Deflate compressed data format specification version 1.3," RFC 1951, May 1996. [Online]. Available: http://www.ietf.org/rfc/rfc1951.txt

[24] K. Sayood, *Introduction to Data Compression*, 3rd ed. San Francisco, CA: Elsevier, 2006.

[25] C. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, Sept. & Oct. 1948.

[26] R. Hamming, *Coding and Information Theory*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1986.

[27] H. Louie, "Characterizing and modeling aggregate wind plant power output in large systems," in *Proc. PES General Meeting*, Minneapolis, MN, Jul. 2010.

[28] ——, "Evaluation of probabilistic models of wind plant power output characteristics," in *Proc. Probablistic Methods Applied to Power Systems*, Singapore, Jun. 2010.

[29] International Electrotechnical Commission, "Power performance measurements of electricity producing wind turbines," 2005. [Online]. Available: 61400-12-1

Dr. Henry Louie (M '03) received the B.S.E.E. degree from Kettering University in 2002, an M.S. degree from the University of Illinois at Urbana-Champaign in 2004 and a Ph.D. in Electrical Engineering from the University of Washington in 2008. From 2007-2008 he worked for 3TIER Environmental Forecast Group, Inc. He is now an Assistant Professor in the Department of Electrical and Computer Engineering at Seattle University. He is a Member-At-Large of the IEEE PES Governing Board.

Agnieszka Miguel received the B.S. and M.S. degrees in electrical engineering from Florida Atlantic University, Boca Raton, FL, in 1992 and 1994, respectively. She received the Ph.D. degree in electrical engineering from the University of Washington, Seattle, in 2001. She is currently an Associate Professor and Chair in the Electrical and Computer Engineering Department at Seattle University. Her research interests include data compression, image processing, and engineering education.