International Journal for Uncertainty Quantification, 1(1):xxx–xxx, 2012

THE MUTUAL INFORMATION DIAGRAM FOR UNCERTAINTY VISUALIZATION

Carlos D. Correa & Peter Lindstrom*

Center for Applied Scientific Computing (CASC), Lawrence Livermore National Laboratory, Livermore, CA,USA

Original Manuscript Submitted: 09/05/2012; Final Draft Received: 09/05/2012

We present a variant of the Taylor diagram, a type of 2D plot that succinctly shows the relationship between two or more random variables based on their variance and correlation. The Taylor diagram has been adopted by the climate and geophysics communities to produce insightful visualizations, e.g., for intercomparison studies. Our variant, which we call the Mutual Information Diagram, represents the relationship between random variables in terms of their entropy and mutual information, and naturally maps well-known statistical quantities to their information-theoretic counterparts. Our new diagram is able to describe non-linear relationships where linear correlation may fail; it allows for categorical and multi-variate data to be compared; and it incorporates the notion of uncertainty, key in the study of large ensembles of data.

KEY WORDS: *mutual information, entropy, variation of information, uncertainty visualization, Taylor diagrams*

1. INTRODUCTION

Making sense of statistical data from observation and simulation is challenging, especially when data is surrounded by uncertainty. Dealing with uncertainty is a multi-faceted process that includes the quantification, modeling, propagation and visualization of the uncertainties and errors intrinsic in the data and arising from data transformations. In this paper, we focus on the visualization of uncertainty.

One common task in uncertainty studies is the comparison of different models to determine whether they are effective in explaining the observed data. In the past, this has been addressed as finding correlations and similarities between data distributions generated from the different models, and quantifying the fit between the model and the observation via statistical analysis. Visualization tools, typically scatter plots or plots of summary statistics (such as means and variances), often restrict the analysis to pairwise comparisons. A full comparison among a large number of models quickly becomes impractical.

To address this issue, Taylor [1] developed a succinct plot that represents several statistics of two random variables simultaneously, now commonly referred to as the Taylor diagram. This plot, quickly adopted by the geophysics and climate communities as the de facto standard for performing intercomparison studies [2], relates three key statistics that describe the relationship between two variables in terms of their variances, their correlation, and their centered root mean square (CRMS) difference.

The Taylor diagram supports important tasks such as measuring the similarity between two variables, understanding the observational uncertainty of a variable of interest, and measuring the improvement of the model skill after refinement, e.g., after increasing the number of samples of the variables or changing the sampling strategy.

However, not all relationships can be explained in terms of variance and correlation alone. In many cases variables exhibit non-linear dependence, a fact that cannot be correctly identified using linear correlation. In other cases, in the presence of outliers, two variables may exhibit low correlation, while otherwise being relatively similar.

Motivated by these limitations, we propose a new type of plot, the *Mutual Information Diagram* (MID), which

*Correspond to: Carlos D. Correa, E-mail: correa23@llnl.gov. Prepared by LLNL under Contract DE-AC52-07NA27344.

2152–5080/10/$35.00

incorporates other types of relationships between two distributions to highlight those similarities and differences not identifiable via statistical quantities. Our approach is to analyze the similarity between two distributions in terms of their shared information. Specifically, our diagram represents the marginal entropies of two distributions, their mutual information, and their *variation of information* [3]—a metric over the space of probability distributions.

This new information-theoretic plot has certain advantages over its statistical predecessor, in that it can expose non-linear relationships between distributions, it is far less sensitive to outliers (e.g., due to noisy measurements or partially corrupt or missing data), and unlike the Taylor diagram it is applicable to both numerical and categorical data.

We show a number of examples where we use the MID to compare different models and make inferences about the sensitivity and uncertainty of multiple variables in an ensemble of simulations. Based on our results, we believe the proposed technique is a useful visualization tool that can be used in conjunction with the traditional Taylor diagram and other graphical representations to highlight the different aspects that relate a collection of distributions.

2. RELATED WORK

Visualizing differences and similarities between models has been a challenge since the generation of the first statistical diagrams. The extension of scatterplots to handle multi-variate relationships has led to metaphors such as the scatterplot matrix and views of parallel coordinates [4]. However, these visualizations seldom scale up to a large number of distributions. Instead, it becomes more effective to simultaneously visualize several distributions. Trend charts plot distributions in a single diagram to reveal correlations. Quartile charts improve trend charts with statistical information, such as the minimum, maximum, percentiles and the median [5]. To improve scalability, a common strategy is to turn to summary measures of the data, such as first and second order statistics. The Boxplot, for example, is a diagram that represents several statistics of a distribution to facilitate visual comparison [6]. Several extensions have been proposed, such as the Violin plot [7], the Beanplot [8], and others suggested by Benjamini [9]. In an attempt to provide a general solution, Potter et al. provide a unified framework for visualizing summary statistics and uncertainty [10]. These plots are primarily designed to highlight statistics of a variable or a distribution, but do not intrinsically encode the relationship between two or more variables, since they are often embedded in the original metric space of the data. Our diagram, on the other hand, maps distributions or variables to points in a 2D metric space where distances encode the variation of information between distributions.

Closest to out work is the Taylor diagram, proposed by Karl Taylor to visualize the correlation and RMS difference among several variables [1]. In his diagram, a random variable is plotted as a point in a 2D plane whose location is given by the standard deviation and correlation with respect to a reference variable. Because of the way this diagram is created, the distance to the reference model in this diagram corresponds to the centered root mean square difference (the RMS difference after subtracting off the mean). In this paper, we propose a new diagram based on information theory instead of first and second order statistics. In our diagram, the location of a model is given by its entropy and its mutual information with a reference distribution.

Mutual Information analysis has been shown to be an effective way of comparing clusterings and studying relationships between time-series, and has been applied successfully in information security and gene expression analysis [11, 12]. Numerous methods have been proposed to estimate both entropy and mutual information, some of which rely on estimating an underlying probability density function or directly estimating these quantities via fitting [13–15]. In this paper, we are not concerned with how to estimate entropy and mutual information, but rather with how to present this information intuitively in a diagram that retains certain metric properties [12, 16].

3. BACKGROUND

3.1 The Taylor Diagram

The Taylor diagram is a graphical representation of the statistical relationship between two random variables. This diagram represents three population statistics relating two variables simultaneously: their standard deviations, their correlation and their centered root mean square difference. Consider two discrete random variables X and Y of n

corresponding values each, with mean μ_X and μ_Y and standard deviation σ_X and σ_Y, respectively. Let R_{XY} denote Pearson's correlation coefficient

$$R_{XY} = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} \tag{1}$$

where

$$\mathrm{cov}(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y) \tag{2}$$

is the covariance between X and Y. The centered RMS difference between X and Y is

$$RMS(X,Y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\big((x_i - \mu_X) - (y_i - \mu_Y)\big)^2} \tag{3}$$

The Taylor diagram exploits the triangle inequality formed by these statistics:

$$\begin{aligned} RMS(X,Y)^2 &= \sigma_X^2 + \sigma_Y^2 - 2\sigma_X\sigma_Y R_{XY} \\ &= \sigma_X^2 + \sigma_Y^2 - 2\,\mathrm{cov}(X,Y) \\ &= \mathrm{cov}(X,X) + \mathrm{cov}(Y,Y) - 2\,\mathrm{cov}(X,Y) \end{aligned} \tag{4}$$

The diagram is a polar plot based on the law of cosines:

$$c^2 = a^2 + b^2 - 2ab\cos\theta \tag{5}$$

The Cartesian coordinates of a variable Y with respect to a reference variable X are thus given by $(\sigma_Y \cos\theta_{XY}, \sigma_Y \sin\theta_{XY})$, with $\theta_{XY} = \cos^{-1} R_{XY}$. These quantities are shown graphically in Fig. 2(1).

Although half Taylor plots, which show only one quadrant, are common, we here compare to the full Taylor diagram, which uses two quadrants to represent both positive and negative correlations.

One of the limitations of the Taylor plot is that it is assumed that the similarities and differences between two distributions can be explained solely using correlation and standard deviation. However, this may not be the case for many distributions. Anscombe highlights this issue to advocate the use of diagrams in lieu of purely statistical reporting of the data [17]. As part of his argument, he devised a simple set of synthetic data sets, all of which have the same correlation and standard deviation, but that are clearly different. This has become known as Anscombe's quartet, three of which are depicted in Fig. 1(1). For these distributions the correlation between any of *B*, *C* and *D* with *A* ($a_i = i$) is about 0.816 and their standard deviation is 1.93. Therefore, all three distributions appear at the exact same position in the Taylor diagram. But clearly the distributions are not the same, nor do they exhibit a similar degree of randomness and dependence on *A*. To better capture this dependence, we resort to information theoretical quantities such as entropy and mutual information, as described below.

3.2 Entropy and Mutual Information

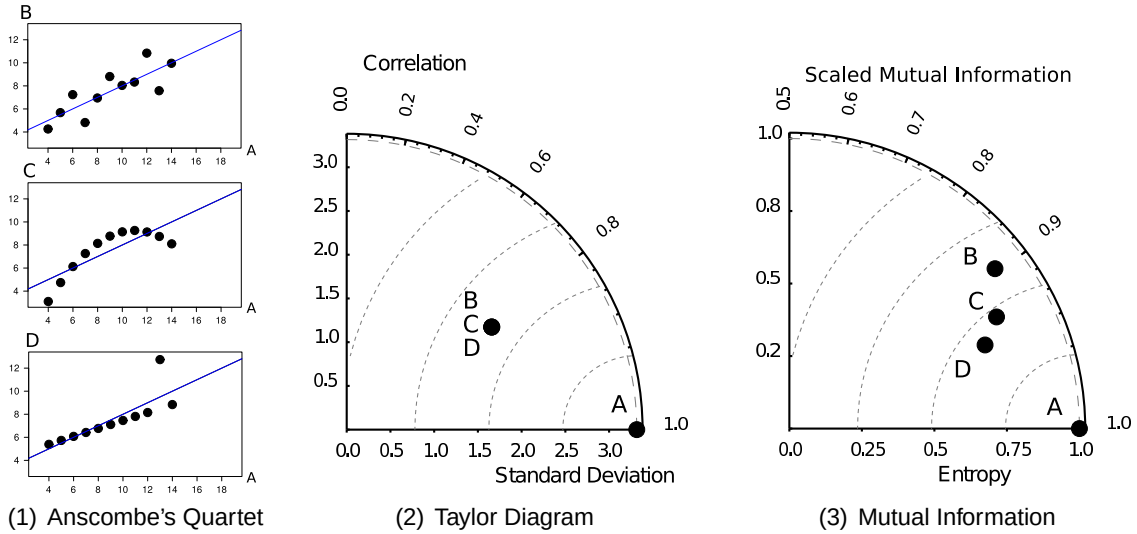
Entropy is a measure of the uncertainty associated with a random variable, or, alternatively, a measure of the amount of information contained in a distribution. For a discrete distribution X, the Shannon entropy is defined as

$$H(X) = -\sum_{i=1}^{n} p(x_i)\log p(x_i) \tag{6}$$

where $p(x_i)$ is the probability of an outcome x_i.

Shannon's entropy applies to discrete distributions. A continuous definition of entropy, also known as *differential entropy*, is defined as

$$h(X) = -\int_X f(x)\log f(x)dx \tag{7}$$



(1) Anscombe's Quartet (2) Taylor Diagram (3) Mutual Information

FIG. 1: (1) Three distributions from Anscombe's quartet [17], all of which have the same correlation and standard deviation but that have a different degree of randomness and dependence on A. (2) In the Taylor diagram, they appear in the same position and we cannot distinguish their dependence on A. (3) In the Mutual Information Diagram, the three distributions appear at different locations and suggest that D shares more information with A than B and C do.

where $f(x)$ denotes the probability density.

Kraskov et al. [13] showed that the discrete and differential entropy can be related via an approximation

$$H(X) \approx h(X) - \log\Delta \tag{8}$$

by considering the limit $\lim_{\Delta\to 0} H(X)$ of discrete entropy, where Δ is the discretization length.

Mutual information (MI) describes the information that two distributions share [13], and is defined as

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)\log\frac{p(x,y)}{p_X(x)p_Y(y)} \tag{9}$$

where $p(x,y)$ is the joint probability of X and Y, and $p_X(x)$ and $p_Y(y)$ are the marginal probabilities of X and Y, respectively. Alternatively, one can express mutual information as

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{10}$$

where $H(X)$ and $H(Y)$ are the marginal entropies and $H(X,Y)$ the joint entropy of X and Y.

When the mutual information of two distributions X and Y is zero, knowing X does not give any information about Y and vice versa, i.e., X and Y are independent. Conversely, since $H(X,X) = H(X)$, the mutual information of two identical distributions equals their entropy.

Estimating entropy and mutual information is challenging, and different methods have been proposed in the past [11, 13]. One of the most commonly used methods estimate the underlying joint and marginal distributions first, using non-parametric models such as histograms or kernel density estimators. Alternatively, one can attempt to estimate these quantities using model fitting, as suggested by Suzuki et al. [15]. In this paper, we compute entropy and mutual information using kernel density estimation (KDE). Later on in Section 4.5, we show how the resulting diagrams differ depending on the parameters used for KDE. Although these estimates may not be optimal, the study of the accuracy of entropy and mutual information estimates is well beyond the scope of this paper.

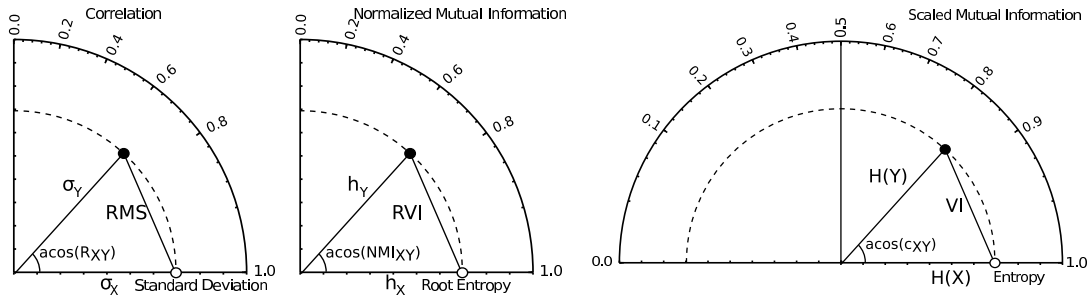


FIG. 2: Relationship between the Taylor diagram and the Mutual Information diagram. (1) The Taylor diagram relates standard deviation σ_Y, correlation R_{XY} and centered root mean square error RMS. (2) Analogously, the RVI-based MI diagram relates root entropy h_Y, normalized mutual information NMI and root variation of information RVI. (3) We may also define the MI diagram in terms of entropies, which relates entropy $H(Y)$, a scaled and biased version of MI and Variation of Information VI. Note that this plot spans two quadrants.

4. MUTUAL INFORMATION DIAGRAM

In this paper, we propose a new type of plot, called the Mutual Information Diagram (MID), which is the analogous version of the traditional Taylor diagram in terms of information and uncertainty. To arrive at such a diagram, we first describe the key relationship between entropy and mutual information.

4.1 Triangle Inequalities Involving Entropy

To extend the traditional Taylor plot, we examine triangle inequalities in information theory over probability distributions. In particular, we make use of the metric known as the *Variation of Information* (VI) between two distributions X and Y [3], defined as:

$$VI(X,Y) = H(X) + H(Y) - 2I(X;Y) = I(X;X) + I(Y;Y) - 2I(X;Y) \tag{11}$$

This quantity is in fact a metric, and satisfies properties such as non-negativity, symmetry, and triangle inequality. Now let us consider the quantify $d(X,Y) = \sqrt{VI(X,Y)}$, which we show is also a metric.

Lemma 1. [ ]
If $d(X,Y)^2$ is a metric, then so is $d(X,Y)$.

Proof. We need to prove four properties:

i. $d(X,Y) \geq 0$. This follows from the definition of d as the principal square root of d^2.

ii. $d(X,Y) = 0 \iff X = Y$. This follows from d^2 being a metric, and thus $d(X,Y)^2 = 0 \iff X = Y$.

iii. $d(X,Y) = d(Y,X)$. This follows from $d(X,Y)^2 = d(Y,X)^2$.

iv. $d(X,Z) \leq d(X,Y) + d(Y,Z)$. Since d^2 is a metric, we have

$$\begin{aligned}(d(X,Y) + d(Y,Z))^2 &= d(X,Y)^2 + d(Y,Z)^2 + 2d(X,Y)d(Y,Z) \\ &\geq d(X,Y)^2 + d(Y,Z)^2 \\ &\geq d(X,Z)^2\end{aligned}$$

As a consequence, since d is non-negative, we have $d(X,Z) \leq d(X,Y) + d(Y,Z)$.

□

Because both d and d^2 are metrics, we may consider diagrams analogous to the Taylor plot involving either metric.

4.2 RVI-Based Diagram

We begin by considering $d = RVI = \sqrt{VI}$ as a metric, and explore the similarities between Eqs. 4 and 11. Let $h_X = \sqrt{H(X)}$ and $h_Y = \sqrt{H(Y)}$. Then

$$\begin{aligned} RVI(X,Y)^2 &= H(X) + H(Y) - 2I(X;Y) \\ &= h_X^2 + h_Y^2 - 2h_X h_Y \frac{I(X;Y)}{h_X h_Y} \end{aligned} \tag{12}$$

This suggests the following mapping between quantities in the Taylor diagram and our new diagram:

- Error $RMS(X,Y) \iff$ root variation of information $RVI(X,Y)$.
- Variance $\sigma_X^2 \iff$ entropy $H(X)$.
- Covariance $\mathrm{cov}(X,Y) \iff$ mutual information $I(X;Y)$.
- Correlation $R_{XY} = \frac{\mathrm{cov}(X,Y)}{\sqrt{\sigma_X^2 \sigma_Y^2}} \iff NMI_{XY} = \frac{I(X;Y)}{\sqrt{H(X)H(Y)}}$.

Here NMI_{XY} denotes the *normalized mutual information* between X and Y [12]. Note also that $\sigma_X^2 = \mathrm{cov}(X,X)$ and $H(X) = I(X;X)$.

Analogous to the traditional Taylor diagram, we build our MI diagram by plotting in polar coordinates the (root) entropy h_Y of distribution Y at an angle defined by $\theta_{XY} = \cos^{-1} NMI_{XY}$.

One aspect to mention about this diagram is that mutual information is non-negative, and thus $0 \leq NMI_{XY} \leq 1$, while the correlation R_{XY} lies in the interval $[-1,1]$. In other words, in this diagram, negative correlations map to positive mutual information. Fig. 2 shows the mapping between the Taylor diagram and the MI diagram.

4.3 VI-Based Diagram

The RVI-based diagram defines distances in terms of the root entropy h_X instead of the actual entropy. In some cases, it becomes easier to read these diagrams if they are defined in terms of entropy and VI. Revisiting Eq. 11, we obtain another relationship by squaring both sides:

$$\begin{aligned} VI(X,Y)^2 &= (H(X) + H(Y) - 2I(X;Y))^2 \\ &= H(X)^2 + H(Y)^2 - 2H(X)H(Y)c_{XY} \end{aligned} \tag{13}$$

with

$$c_{XY} = 2I(X;Y)\frac{H(X,Y)}{H(X)H(Y)} - 1 \tag{14}$$

The quantity c_{XY} is a biased and scaled version of the mutual information. Unlike the RVI-based diagram, the VI-based diagram measures a scaled mutual information:

$$SMI_{XY} = I(X;Y)\frac{H(X,Y)}{H(X)H(Y)} \tag{15}$$

Similarly, we now define the coordinates of a point using the entropy $H(Y)$ and an angle defined by $\theta_{XY} = \cos^{-1} c_{XY}$. This new plot places c_{XY} in the range $[-1,1]$. Fig. 2(3) shows the plot and the corresponding quantities.

4.4 Properties

Compared to the traditional Taylor diagram, based on covariance, the MI diagram based on mutual information has certain advantages:

i. Mutual information is able to capture both linear and non-linear correlations.

ii. Our new diagram is not limited to numerical data, but can also be used to plot categorical data, e.g., to show similarities in clusterings, classifications and symbolic sequences.

iii. Entropy and mutual information are rather insensitive to outliers. Even a single outlier can arbitrarily impact both the variance and correlation between two distributions, thus obscuring the similarity of two closely related variables.

iv. Although not explored in this paper, our framework has the potential to be extended to show correlations between multivariate distributions (using multivariate MI), as well as between combinations of two or more distributions and a reference distribution (using co-information), e.g., to explain interactions between multiple variables.

4.5 Entropy Estimation

Unlike the Taylor diagram, which uses statistics that can be directly derived from the discrete sample data, MI diagrams require an estimation of the underlying probability density functions [18]. Alternatively, one can directly estimate the mutual information and entropies without explicitly estimating the underlying density [13]. It is not the focus of the paper to propose a new or suggest a preferred method for estimating the mutual information, so we adopt techniques found in the literature, some of which estimate the underlying probability density first.

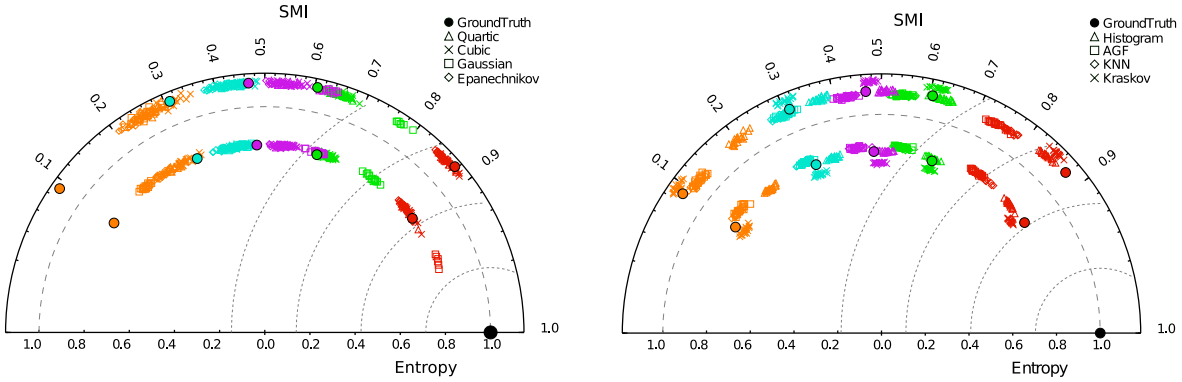
To demonstrate how the choice of the method affects the diagram, we have computed the mutual information and entropy of a number of bivariate normal distributions G_{ij} with marginal probability distributions $X \sim N(0,1)$, $Y_{ij} \sim N(0,\sigma_i)$ and correlation R_j with respect to X, for $\sigma_i \in \{0.5, 1.5\}$ and $R_j \in \{0.5, 0.8, 0.9, 0.95, 0.99\}$ for a number of methods. These include four methods that estimate the probabilities using histograms, adaptive gaussian filtering (AGF) [19], k nearest neighbors, and kernel density estimation (KDE), as well as a method by Kraskov et al. [13], which estimates mutual information and entropy directly. The differential joint entropy of these variables is given by

$$h(X,Y_{ij}) = \frac{1}{2}\ln\big((2\pi e)^2\sigma_i^2(1-R_j^2)\big) \tag{16}$$

We can then approximate the discrete joint and marginal entropies using Eq. 8.

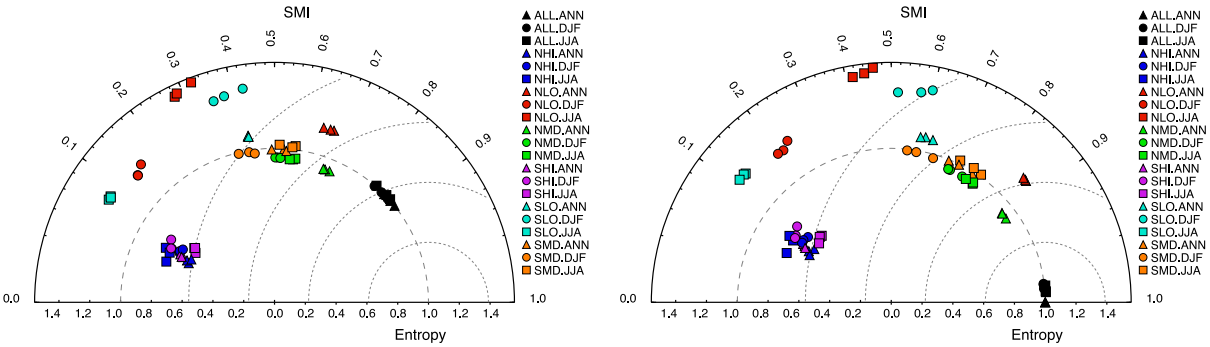
Fig. 3(2) compares four of these methods, where the solid circles represent the ground truth entropy and mutual information for the different 2D variables (color denotes the correlation R_j: 0.5 (orange), 0.8 (teal), 0.9 (purple), 0.95 (green) and 0.99 (red)). For each of the methods, we obtained twenty random draws of 2000 points each. We see that there is an overall agreement of the different methods. Density based methods (histogram, AGF and KNN) estimate the entropies accurately (radial coordinate), but have a noticeable discrepancy in the mutual information (angular coordinate). On the other hand, Kraskov's method has higher accuracy when estimating MI, but larger error when estimating the marginal entropy $H(Y)$. Fig. 3(1) compares the methods based on kernel density estimation for different choices of kernel, namely quartic, cubic, Gaussian and Epanechnikov, using 40 bins and an estimate of the optimal bandwidth [20]. Compared to the previous methods, all these are good estimators of the marginal entropies, but have noticeable bias in the estimation of mutual information. Nonetheless, the behavior of each estimator is consistent for the different variables. This suggests that, even though the MI diagram is sensitive to the choice of method and parameters in estimating the entropies, the relative location of the distributions in the diagrams is more or less preserved.

It must be noted that for some of these methods, in particular for KDE methods, it is not always the case that $I(X;X) = H(X)$. This implies that the reference variable does not necessarily appear on the x-axis of the diagram. To counteract this, we normalize the data so that the mutual information $I(X;Y)$ is defined relative to the mutual



(1) KDE Methods (2) Other Estimation Methods

FIG. 3: Comparison of different mutual information and entropy estimators for a number of random draws of bivariate normal distributions with marginal standard deviation $\sigma_X = 1.0$ and $\sigma_Y \in \{0.5, 1.5\}$ and correlations $R_{XY} \in \{0.5, 0.8, 0.90, 0.95, 0.99\}$. (1) Kernel density estimators with fixed bandwidth and different choice of kernel. (2) Other estimation methods, including histograms, adaptive Gaussian kernels, k nearest neighbors and Kraskov's method [13] The direct method by Kraskov et al. yields a better MI estimate than density based methods. Compared to (2), KDE-based methods have a larger error in MI. Despite the evident errors, the relative location of these variables in the diagram is more or less preserved for each individual choice of estimator.



(1) Before Scaling (2) After Scaling

FIG. 4: Mutual Information diagrams before and after scaling for the analysis of climate ensembles. Each variable represents a temporal and zonal average of average precipitation, for three different sets of ensembles. Scaling ensures that the reference variable, in this case the annual and global average, always lies on the x-axis. Although similar, some ensembles that seem identical before scaling (e.g., SLO.ANN, SLO.JJA) are actually different (with respect to the reference variable) when the scaling is applied.

information $I(X;X)$. We achieve this by simply scaling the mutual information, so that our diagrams make use of the scaled mutual information:

$$I(X;Y) = \tilde{I}(X;Y)\frac{H(X)}{\tilde{I}(X;X)} \tag{17}$$

where $\tilde{I}(X;Y)$ is the *estimated* mutual information using one of the methods discussed above. We show the effect of scaling in Fig. 4.

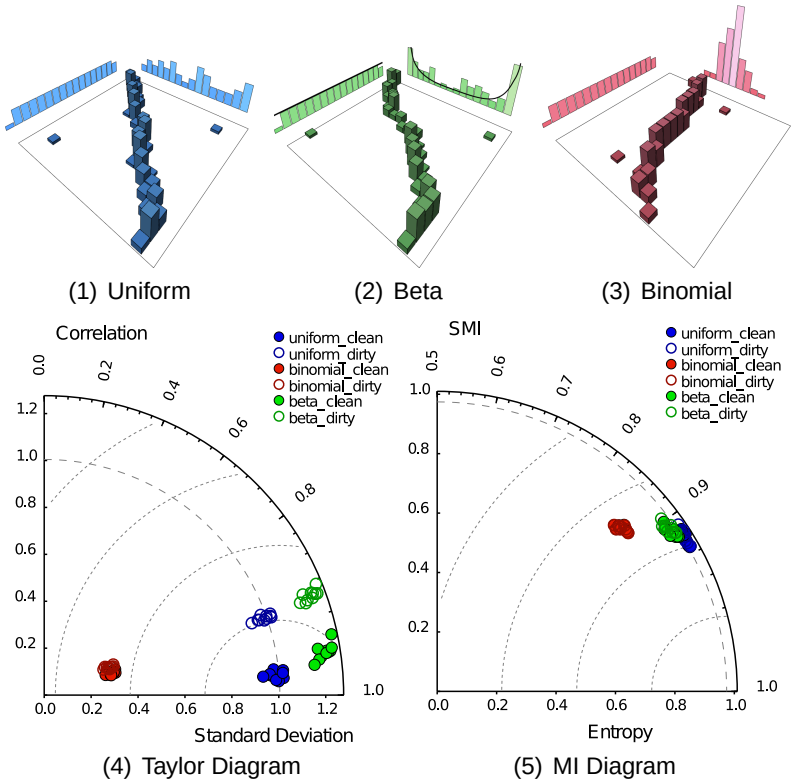


FIG. 5: Resilience to outliers of the MI diagram. We show the Taylor plot for different random draws from three distributions: uniform (1), binomial (2) and beta (3) The reference distribution is uniform. When adding outliers (4), these draws appear in different locations in the Taylor diagram, since outliers affect the correlation (notice that, due to the design of this experiment, the standard deviation is not affected). A MI diagram (5), on the other hand, is resilient to outliers, and draws from the same distribution appear roughly in the same location in the presence or absence of outliers.

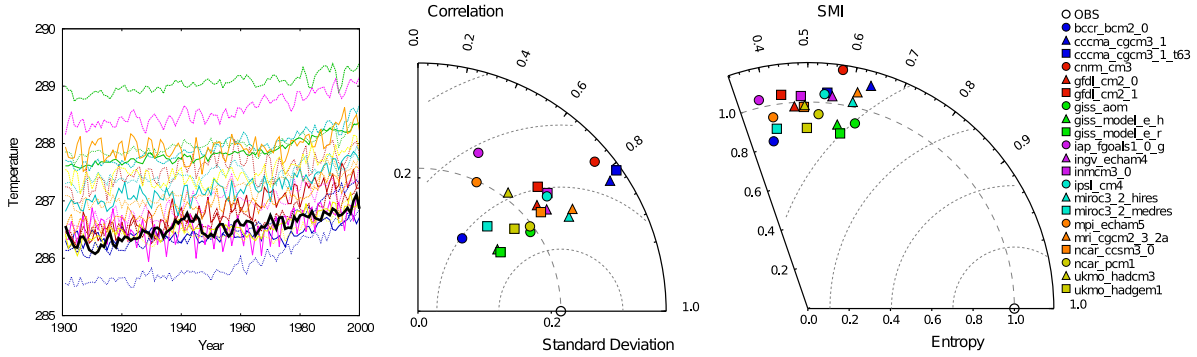
4.6 Effects of Outliers

An important aspect of mutual information is its relative resilience to the presence of outliers compared to correlation. To illustrate this property, we have created a number of discrete data sets ($n = 128$) from three different distributions — uniform, binomial and beta (with $\alpha = \beta = \frac{1}{2}$). We introduce outliers by interchanging two samples at random, which does not change the marginal distributions. The resulting distributions and the near-uniform reference distribution are depicted in Figs. 5(1-3), as 2D histograms, together with their respective marginal distributions. Although this interchange does not affect the variance or entropy, we expect the correlation to be more sensitive than the mutual information to such a perturbation.

Figs. 5(4) and 5(5) show the resulting Taylor and MI diagrams. Notice how, when using mutual information, the corresponding distributions cluster together regardless of whether outliers are present or not. In the Taylor diagram, we see a clear separation between the distributions generated with outliers (dirty) and those without (clean).

5. RESULTS

We have explored our new diagram in a series of applications, from comparing climate models and different ensemble runs for uncertainty quantification, to the evaluation of clustering algorithms.



(1) Temperature (2) Taylor Diagram (3) MI Diagram

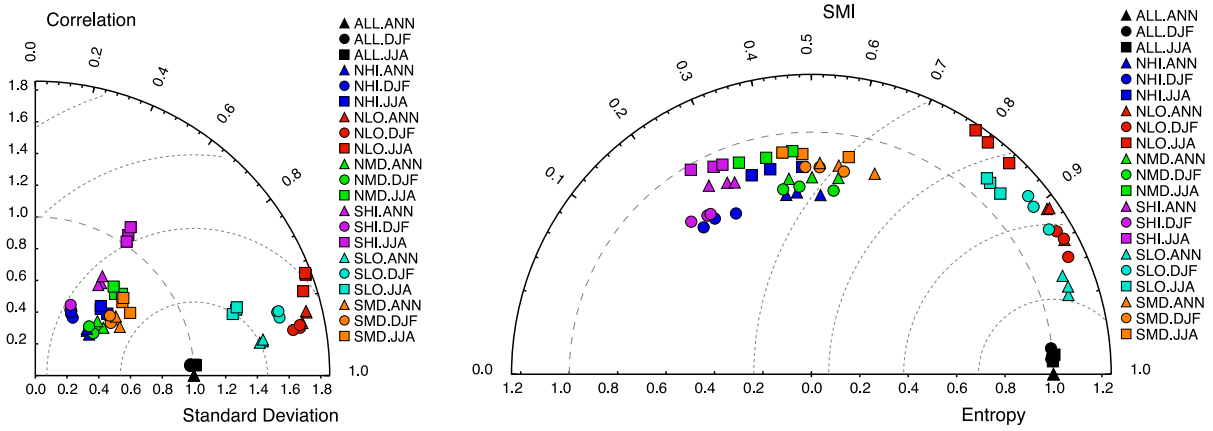
FIG. 6: Intercomparison study of merged surface air temperature from CMIP3 for the 20th century scenario. In these diagrams, each data point corresponds to the time series of annual average in temperature during the 20th century from models obtained at different climate centers around the world. (1) shows the time series in a trend plot. (2) The Taylor diagram with respect to observations. (3) MI diagram. Notice how the differences between the two diagrams may help understand the different distributions. For example, we see that the cccma_cgcm3 models (blue square and triangle) have similar correlations and standard deviation. However, in terms of information, one of them has a higher mutual information with the observation, possibly suggesting a non-linear relationship between the two.

5.1 Intercomparison Studies

As part of a global initiative to understand climate change, the Program for Climate Model Diagnosis and Intercomparison (PCMDI) has collected model output of climate simulations from leading modeling centers around the world. One of the data sets in this collection, called the Coupled Model Intercomparison Project (CMIP), studies output from coupled ocean-atmosphere general circulation models that also include interactive sea ice, and includes realistic scenarios, such as the twentieth century scenario, the preindustrial control scenario and a scenario that projects 300 years into the future [2].

Here, we study the relationship between the different models for the annual average of merged surface air temperature in the 20th century. This data set consists of 21 models from centers around the world. Understanding the differences and similarities of the various models helps us quantify the uncertainty of climate simulations and increase our confidence in conclusions regarding climate change.

Fig. 6(1) shows the average annual temperature time series for each model and the observed temperature. Naturally, it becomes difficult to make sense of the plot and suggest which models are more similar to the observation. A pairwise diagram, such as a scatterplot matrix, becomes too large for effective visual exploration. Fig. 6 compares the resulting Taylor and Mutual Information diagrams with respect to observations, which work better as global views. Mutual information and entropy are computed by estimating the underlying PDF using a histogram. Although there seems to be agreement of which models differ most from the observations (for example, the iap_fgoals model, purple circle), a few models are placed differently relative to each other in the two diagrams, suggesting non-linear relationships between the different models. For example, the cccma_cgcm3_t63 model (blue square), appears as similar to the other cccma model (blue triangle) in terms of correlation, but not so much in terms of mutual information. As we will see in the following example, this would not be the case if these distributions were approximately normal. With the two diagrams, studying these differences may help identify more accurately the similarities and dissimilarities between models.



(1) Taylor diagram (2) Mutual Information diagram

FIG. 7: Uncertainty quantification of climate models. In these diagrams we compare several output variables of a climate simulation involving FLUT (upwelling long wave flux at the top of the atmosphere). These variables are formed by seasonal and zonal averages of FLUT. The seasonal averages (shape) represent the annual flux during summer months (in the northern hemisphere – JJA) and winter months (DJF) and the zonal averages (color) comprise three latitudinal zones in the northern (NHI, NMD, NLO) and southern (SHI, SMD, SLO) hemispheres. For each of these, we plot the output for three ensembles, containing 561, 895 and 1318 runs each. Although the two diagrams are different, we notice that relative locations are similar and one diagram can be explained (mostly) as a warping of the other. Nonetheless, we are able to stress the differences more easily in the MI diagram, which suggests similar entropies along each of the sets, but different degrees of mutual information. Note how, while in the Taylor diagram at least two of the ensembles agree (see how some pairs of figures of same color and shape, corresponding to the same seasonal and local average but using different ensembles, overlap), they are more clearly separated in the MI diagram.

5.2 Analysis of climate ensembles

Another application deals with the comparison of a number of ensemble runs to quantify uncertainty and sensitivity in climate simulations. To this end, climate scientists have identified a subset of 27 input parameters of the Community Atmosphere Model (CAM), and are studying the impact of perturbations of these parameters in climate-related outputs, such as the upwelling long wave flux at the top of the atmosphere (FLUT) and total precipitation rate (PRECT).

To explore this high dimensional space, simulations are run using different parameter sampling strategies, such as latin-hypercube and one-at-a-time sampling. Fig. 4 shows the MI diagram for PRECT, the total precipitation rate, for different annual and seasonal averages (DJF for Dec-Jan-Feb months, JJA for Jun-Jul-Aug months), and zones (Northern/Southern HIgh, MeDium and LOw zones). For each of these, we compare three different sets of ensemble runs, containing 561, 896 and 1318 runs each. Fig. 7 depicts the Taylor and MI diagrams for FLUT. To estimate entropy, we compute the probability density function via kernel density estimation using the optimal bandwidth for a bivariate normal distribution and Epanechnikov kernels. We see that in both the Taylor and MI diagrams, the relative locations of the data points are fairly well preserved, i.e., knowing one of the diagrams explains to some extent the other diagram. This is not too surprising, since these ensembles are obtained from annual and/or zonal averages, and the resulting distributions are approximately Gaussian, for which correlation and mutual information tell a similar story. Nonetheless, we observed both for PRECT and FLUT that the differences within each zone and season are more pronounced in the MI diagram. Specifically, we notice that the first and last set, containing 561 and 1319 runs, respectively, differ consistently from the second set. In fact, the second set differs from the others in that it was obtained with *one-at-a-time* (OAT) and *Morris-one-at-a-time* (MOAT) sampling strategies, where one input parameter is changed between consecutive runs [21].

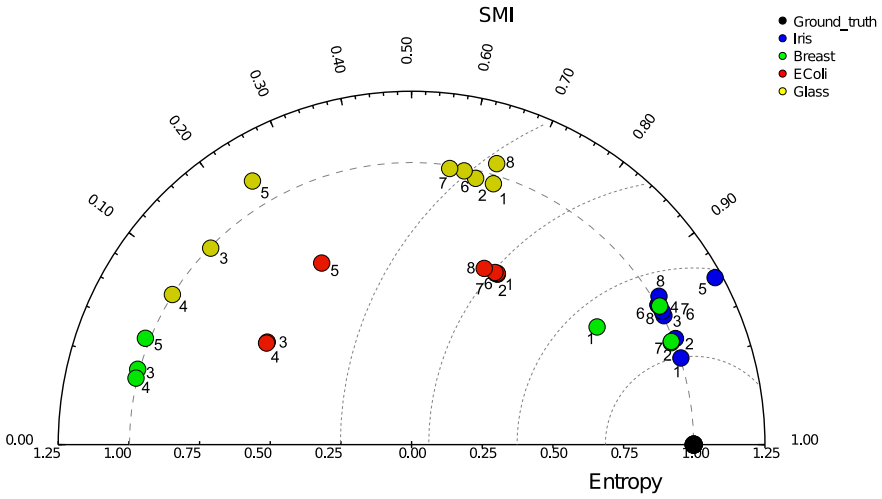


FIG. 8: A MI diagram that summarizes a study comparing several kernel clustering algorithms for four data sets: Iris (blue), Breast (green), E-coli (red) and Glass (yellow) [22]. The algorithms compared are: 1:FCM-I-fs, 2:FCM-II-fs, 3:FCM-I-km, 4:FCM-II-km, 5:SVC, 6:FCM-I, 7:FCM-II, 8:Kmeans. In agreement with the authors' own study, we see that no single method appears to outperform the others, but the diagram clearly suggests that feature space clustering consistently performs well. In contrast, we have a better sense of how poorly methods that incorporate kernelization (3,4) and SVC (5) behave in relation to the other methods and the ground truth (where MI is practically zero). This diagram proves to be an essential tool to summarize what otherwise has required table comparisons.

5.3 Clustering and Classification

As mentioned above, unlike the Taylor diagram, our MI diagram can also be applied to depict similarities between symbolic data. An important application of this is the study of classification and clustering algorithms. Mutual information (and VI) has become a practical metric for measuring the performance of clustering algorithms. In general, once a ground truth classification is known, the *confusion (or matching) matrix* defines the joint probability associated with two instances of the classification data. Thus, given the set of classes X and predicted clusters Y, the probability $P(x_i, y_j) = N_{i,j}/N$, where $N_{i,j}$ is the number of responses of the confusion matrix at position i, j and N is the number of elements in the classification data. Since we start from discrete distributions, we can directly compute entropy and mutual information without the need for an estimation step.

Our MI diagram thus provides a single view to compare and make inferences about the performance of different clustering methods. We applied our technique to the results of a comparison study of several kernel clustering algorithms [22]. This study consisted of five classification problems (of which we depict four), and compared the result of eight different algorithms: 1:FCM-I-fs, 2:FCM-II-fs, 3:FCM-I-km, 4:FCM-II-km, 5:SVC, 6:FCM-I, 7:FCM-II, 8:Kmeans. FCM refers to Fuzzy c-means methods, with two different objective functions, one being an L-2 functional (I) and the other introducing a maximum entropy criterion (II). These are algorithms 6 and 7. In addition, two variants of these are included in the study, one clustering in feature space (1,2) and the other using a kernelization method (3,4). To complete the study, the authors compare to Support Vector clustering and *k*-means [22]. Fig. 8 shows the normalized MI diagram (i.e., entropies are normalized so that the entropy of the reference variable is 1) that summarizes the results of their study for four data sets: Iris (blue), Breast (green), E-coli (red) and Glass (yellow).

Another example is depicted in Fig. 9, where we decompose a bivariate scalar function into homeomorphic contours (aka. contour tree segmentation). In this study, reported by Correa and Lindstrom [23], the authors show that the choice of neighborhood graph to connect sparse samples of this function has an effect on the quality of the decomposition. Here, we employ the MI diagram to graphically summarize the results reported in this study, which concludes

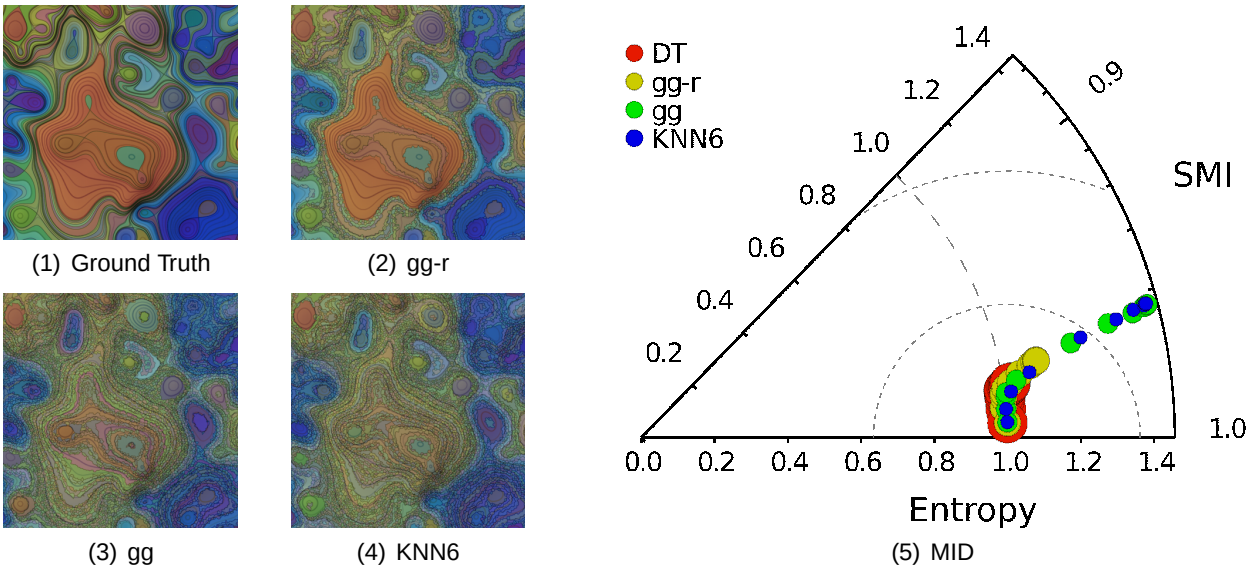


FIG. 9: Evaluating contour tree segmentations of a 2D terrain. Similar to the clustering problem, we evaluated the performance of different neighborhood graphs [23] in estimating the topology of a 2D terrain and computing its contour tree decomposition. The graphs are the Delaunay Triangulation (DT), the Gabriel graph (gg), the relaxed Gabriel graph (gg-r) and *k*-nearest neighbor graph (KNN6). The insets on top show (left to right) the true decomposition of the terrain, the one provided by the relaxed Gabriel graph (also similar to dt—not shown), and the one given by the Gabriel graph (also similar to knn—not shown). We see that the knn6 clustering is much noisier and we should expect a higher entropy, as confirmed by our MI diagram. Each point in the diagram depicts the mutual information and entropy from the resulting decomposition compared to the ground truth decomposition, for different levels of noise reduction. We see that dt and gg-r are considerably better than the other two and more resilient to noise.

that neighborhood graphs such as the Delaunay Triangulation (DT) and the Relaxed Gabriel graph (gg-r) are more accurate and resilient to noise than other graphs, such as the Gabriel graph (gg) and *k*-nearest neighbors (KNN6). Figs. 9(1), 9(2) and 9(4) show the ground truth contour tree segmentation (reference) and the segmentation for the relaxed Gabriel and Gabriel graphs, respectively. The graphs for the Delaunay triangulation (omitted) and KNN are very similar to Fig. 9(2) and Fig. 9(4), respectively. The diagram shows these results for a sequence of decompositions corresponding to various levels of topological denoising. For lower noise thresholds, graphs such as gg and knn result in noisier decompositions and increase the entropy with respect to the ground truth decomposition computed on a dense regular grid. The gg-r graph is much less sensitive to the noise level, and only marginally increases the entropy. On the other hand, it is clear that the Delaunay triangulation does not increase the entropy, and should result in better contour tree segmentations. The MI diagram succinctly condenses the results of the study and allows us to make inferences about the information shared by any given pair of segmentations.

6. LIMITATIONS

Despite the many advantages of mutual information over correlation, there are some limitations of our proposed diagram. First, unlike correlation, which can be computed easily for random variates, estimation methods for entropy and mutual information are more elaborate. A common approach is to approximate the joint and marginal probability functions via histograms or kernel density estimates, but this implies setting parameters such as the bin size, the bandwidth and choice of kernel. The choice of these parameters has a global effect on the resulting MI diagram. Although in our observations, most of the relative locations are preserved, a more informative MI diagram should

represent the envelope of possible locations based on multiple entropy and MI estimation methods. Second, in our plot there is no distinction between negative and positive correlations. In certain applications, this may be important. We believe that a combination of the Taylor diagram and the MI diagram is more effective for discovering important relationships between distributions, such as non-linearities. Alternatively, one can incorporate correlation and standard deviation in the MI diagram with additional graphical attributes such as color or texture.

7. SUMMARY

We have presented the Mutual Information diagram, a novel diagram that relates three key information-theoretic measures: entropy, mutual information and variation of information. In our preliminary studies and feedback from climate scientists, we have identified several use cases for this diagram. (1) Similar to the Taylor diagram, our MI diagram proves to be crucial in understanding how certain output variables from different simulations and models differ from a reference model or observations. Trend lines and scatterplots often obscure statistical or information differences that this plot can succinctly represent in a metric space. (2) When exploring multi-output data, one can generate MI diagrams for the different variables, each using a different variable as a reference. Together, these diagrams encode the relationships between different variables (in terms of dependence and mutual information) that may be too cumbersome to visualize using more traditional scatterplot matrices. (3) Our diagram directly provides a space to visualize the quality of density estimation methods, useful for information-theoretic measures such as entropy and information, as shown in Fig. 3. This diagram helps us quantify not only the uncertainty inherent in the data, but also the added uncertainty when estimating probability density functions. (4) A variety of applications readily compile information-theoretic quantities to measure the quality of their algorithms or visualization parameters. We have shown an example that uses the MI diagram to visualize the accuracy of clustering methods, but it can be further extended to evaluate techniques that use entropy in visualization, such as automatic view selection for volume visualization [24].

There are numerous challenges involved in ensuring an accurate estimate of information-theoretic quantities for different domains, and our novel diagram not only benefits from progress in this area, but also becomes a useful plot for evaluating the different methods.

ACKNOWLEDGMENTS

This work performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. LLNL-JRNL-558392.

REFERENCES

1. Taylor, K. E., Summarizing multiple aspects of model performance in a single diagram, *Journal of Geophysical Research*, 106(D7):7183–7192, April 2001.
2. Meehl, G. A., Covey, C., Delworth, T., Latif, M., McAvaney, B., Mitchell, J. F. B., Stouffer, R. J., and Taylor, K. E., The WCRP CMIP3 multimodel dataset: A new era in climate change research, *Bulletin of the American Meteorological Society*, 88(9):1383–1394, 2007.
3. Meila, M. Comparing clusterings by the variation of information. In: Schlkopf, B. and Warmuth, M. (Eds.), *Learning Theory and Kernel Machines*, Vol. 2777 of Lecture Notes in Computer Science, pp. 173–187. Springer Berlin / Heidelberg, 2003.
4. Inselberg, A., The plane with parallel coordinates, *The Visual Computer*, 1(2):69–91, August 1985.
5. Potter, K., Wilson, A., Bremer, P.-T., Williams, D., Doutriaux, C., Pascucci, V., and Johnson, C. R., Ensemble-Vis: A framework for the statistical visualization of ensemble data, In *IEEE International Conference on Data Mining Workshops*, pp. 233–240, 2009.
6. Williamson, D. F., Parker, R. A., and Kendrick, J. S., The box plot: A simple visual method to interpret data., *Annals of Internal Medicine*, 110(11):916–921, 1989.

7. Hintze, J. L. and Nelson, R. D., Violin plots: A box plot-density trace synergism, *The American Statistician*, 52(2):181–184, 1998.
8. Kampstra, P., Beanplot: A boxplot alternative for visual comparison of distributions, *Journal of Statistical Software, Code Snippets*, 28(1):1–9, 2008.
9. Benjamini, Y., Opening the box of a boxplot, *The American Statistician*, 42(4):257–262, 1988.
10. Potter, K., Kniss, J., Riesenfeld, R., and Johnson, C. R., Visualizing summary statistics and uncertainty, *Computer Graphics Forum*, 29(3):823–832, 2010.
11. Cover, T. M. and Thomas, J. A., *Elements of information theory*, Wiley-Interscience, New York, NY, USA, 1991.
12. Strehl, A. and Ghosh, J., Cluster ensembles—a knowledge reuse framework for combining multiple partitions, *Journal of Machine Learning Research*, 3(3):583–617, March 2003.
13. Kraskov, A., Stögbauer, H., and Grassberger, P., Estimating mutual information, *Physical Review E, Statistical, nonlinear, and soft matter physics*, 69(6):066138, June 2004.
14. Paninski, L., Estimation of entropy and mutual information, *Neural Computation*, 15(6):1191–1253, June 2003.
15. Suzuki, T., Sugiyama, M., Kanamori, T., and Sese, J., Mutual information estimation reveals global associations between stimuli and biological processes, *BMC Bioinformatics*, 10(S-1):S52, 2009.
16. Kraskov, A., Stögbauer, H., Andrzejak, R. G., and Grassberger, P., Hierarchical clustering using mutual information, *Europhysics Letters*, 70(2):278–284, 2005.
17. Anscombe, F. J., Graphs in statistical analysis, *The American Statistician*, 27(1):17–21, February 1973.
18. Silverman, B. W., *Density Estimation for Statistics and Data Analysis*, Chapman & Hall, London, 1986.
19. Mills, P., Efficient statistical classification of satellite measurements, *International Journal of Remote Sensing*, 32(21):6109–6132, 2011.
20. Scott, D. W., *Multivariate Density Estimation: Theory, Practice, and Visualization (Wiley Series in Probability and Statistics)*, Wiley, September 1992.
21. Morris, M. D., Factorial sampling plans for preliminary computational experiments, *Technometrics*, 33(2):161–174, April 1991.
22. Filippone, M., Masulli, F., and Rovetta, S., An experimental comparison of kernel clustering methods, In *Proceeding of the 2009 conference on New Directions in Neural Networks: 18th Italian Workshop on Neural Networks: WIRN 2008*, pp. 118–126, 2009.
23. Correa, C. D. and Lindstrom, P., Towards robust topology of sparsely sampled data, *IEEE Transactions on Visualization and Computer Graphics*, 17(12):1852–1861, 2011.
24. Bordoloi, U. and Shen, H.-W., View selection for volume rendering, In *IEEE Visualization*, pp. 487–494. IEEE Computer Society, Oct. 2005.