To assist both authors and reviewers, submissions documenting the use of statistical (including machine learning (ML)) modeling are required include the following information. **Such submissions that do not address these points will be deemed "premature" submissions and will not be sent for peer review.** Authors may use their own discretion as to whether the documentation of each point of the rubric belongs in the main text, or if some points are better suited for an appendix or online supplement. Authors are referred to the NIST Engineering Statistics Handbook if they are unfamiliar with terminology used in this author guide: https://www.itl.nist.gov/div898/handbook/index.htm

1) **Delineate the use of multiple statistical models.** Often, two (or more) statistical model(s) may be used in statistical modeling research. For example, one model may be used to screen a larger dataset to select and/or engineer the factors and responses that are used for a different/derived dataset for the final ML model. Or the authors may be comparing multiple statistical/ML modeling approaches. The following points in this rubric must be clearly addressed as applicable for each statistical model that is essential to the impact and conclusions of the research.

2) **Provide a clear definition of the process model formulation.** What is the purpose of the model (estimation, prediction, screening, hypothesis testing, regression vs. classification, etc.)? What are the factors (inputs)? What are the co-factors? What are the responses (outputs)? What are the levels? If ML modeling, is it supervised vs. unsupervised vs. reinforcement learning?

3) **Provide a clear definition of the dataset as it is used for statistical modeling.** Can the scientific community access the dataset for future work and/or to verify the results (not a requirement, but strongly encouraged)? If so, how? We strongly recommend that authors include a separate "Methods" subsection (and if applicable an accompanying appendix/supplement/Data in Brief) that clearly documents the dataset used for the ML modeling itself. Usually, it is different from the data that is directly acquired from experiments and/or simulations.

4) **Provide evidence that the dataset supports statistical modeling.** Statistical modeling (including ML), like finite element simulations or electron microscopy, provides a set of tools to help solve or understand a problem, but not every problem. The samples must feasibly represent a random sampling from probability distributions of the factors and responses. The factors and responses must be statistically related. These two assumptions are essential to all statistical models, including ML. Authors must document that the samples used for modeling meet these fundamental assumptions.

5) **Provide evidence that ML is necessary (if an ML submission).** Once it is determined that the data of interest exhibit statistics that warrant statistical modeling, it should be proved that ML is needed, i.e., that some of the interrelations of factors and responses are of higher dimension than can be estimated with simpler statistical models and distributions. One of the fundamental principles of statistical modeling is that the best statistical model is the simplest one that sufficiently describes the factor-response relationships of interest. **Submissions that seem to use arbitrarily complex ML models without first demonstrating that they are necessary will be deemed out of scope.**

6) **Document how the basic statistical properties of the dataset motivate the choice and selection of the statistical modeling approach.** Is there adequate reasoning/rigor provided in stating why the algorithms that are being considered are the best choices to evaluate for this specific dataset or problem? If there is not an obvious scientific/mathematical reason to choose one vs. another, have the authors evaluated all feasible algorithms for their problem and used appropriate methods and metrics to select the best performer(s)? Documentation of the number of samples vs. the number of parameters to be estimated for each ML framework is required. Just as Gauss elimination cannot ever be used to determine 5 unknowns from 3 linearly independent equations, the samples used for estimating the parameters of an ML model must be great enough in number **AND** exhibit enough statistical variation for the use of the ML approach to be feasible. The samples and their statistical variations should have been well described in point 2. Here, those statistics must relate to the selection of ML approach(es).

7) **Document and discuss the parameter estimation (training) of the model.** Was the dataset divided into training vs. test data in a statistically meaningful/justified way? Did the authors properly test their models? Is the bias-variance tradeoff documented for the parameter estimation of the model(s)? Is the model robust to both extrinsic vs. intrinsic error (i.e., overfitting and underfitting)? Did the training address the implicit limitations of the models, such as the independence condition in Naïve Bayes or gradient collapse in convolutional neural networks, and the impact these limitations have on the applicability of a particular model? Were proper model optimization steps taken in tuning all hyperparameters? Are the full range of hyperparameters reported? Does the article provide sufficient instruction for others to identify the decision points and considerations necessary to construct a similar class of model on an unrelated data set? Can those models be replicated by others? Are the models available to peer reviewers/others wanting to verify/build upon this work (not required, but strongly encouraged and likely to increase impact and citations)? If so, has the code been packaged properly to ensure compatible versions of the necessary packages, e.g., tensor flow, keras, scikit-learn, numpy, scipy, Matlab, etc.?

8) **Completely assess the statistical model prediction performances.** Are both bias and variance of the predictions considered in the uncertainty quantification? Are more than just mean value error assessment(s) or R^2 analysis provided for regression models? Are the residuals well behaved (random, centered about 0, and normal distributed)? Are classification model performances well documented to include accuracy, sensitivity, precision, recall, fallout, etc.? What are the possible implications/artifacts of the cofactors? What are the edge cases and how have they been tested? Include documentation/discussion of the proper context for using and interpreting the model.

9) **Discuss scientific and/or engineering impact gained from the use of statistical modeling?** Did the models solve the problem at hand? Did the models perform better than random guessing and/or the previous state of the art? Does the performance of the models provide evident scientific or engineering impact? How was this impact quantified? **Submissions that do not show some future benefit/purpose of the statistical modeling research will not be peer reviewed.**