Foreword to the CERC report ADMLC/2026/1.1 “A review of methods used to assess the performance of atmospheric dispersion models”

The purpose of this foreword is to provide context to the report on atmospheric dispersion model evaluation produced by CERC for the ADMLC (available to download from https://admlc.com/publications/). The foreword also presents a summary of the discussions from the joint CERC and ADMLC workshop on 11 June 2026, where CERC presented their main findings.

Context

In 2019, the representative on the ADMLC committee from the Scottish Environment Protection Agency (SEPA) suggested to the committee that we fund a review of dispersion model evaluation metrics. This was motivated by SEPA’s interest in evaluating dispersion models for application to air quality in urban areas. The idea was supported by other members of the committee and plans were discussed in several subsequent ADMLC committee meetings, during and after the COVID pandemic. The scope of the work widened over this time to include a detailed literature review of model evaluation metrics and their application to several case studies. The ADMLC put the project out to tender in 2023, with the successful bid being awarded to CERC.

Choice of case studies

As part of CERC’s work, they sought the advice of the ADMLC committee members on the choice of suitable case studies. One of the challenges in selecting these case studies was that they needed to include *both* atmospheric dispersion measurements and model predictions, ideally from several models. A second challenge was the diversity of the ADMLC committee members’ interests, including dispersion model applications in air quality, regulatory assessments of chemical and nuclear sites, emergency response, and topics such as volcanic ash transport in the upper atmosphere. This resulted in a wide range of case studies being selected for CERC’s analysis, including data from field tests, wind tunnel experiments, routine air quality monitoring and long-range tracer experiments.

CERC conducted an extensive literature review and identified several dozen model evaluation metrics. The majority of these metrics were then applied to the case studies. As might be expected, there was a range in values of the metrics for different model predictions across different datasets. CERC presented these results and identified two sets of metric threshold values that could be used to classify model performance: one that was based on approximately 50% of the model predictions falling within the thresholds, and a less stringent threshold based on approximately 75% of the model predictions falling within the thresholds.

Model evaluation criteria and threshold terminology

The use of metric threshold values to assess model performance is a well-established concept, and the criteria proposed by Steven Hanna and Joe Chang [1, 2] are often cited. An example of one of their criteria is FAC2 > 0.5, i.e., a model should give predictions within a factor of two of the measurements for more than half the time. One of the notable findings of the CERC study was that their two sets of threshold values (from applying the 50% and 75% criteria) were similar to those published previously by Hanna and Chang, despite them being derived using an alternative methodology from quite different datasets.

The term “threshold” is often associated with regulatory requirements, such as acceptable risk thresholds or harm thresholds. However, within the context of the current work, “threshold” is not used in this sense but instead it simply refers to a yardstick or indicator that can be helpful in assessing if a model's performance is generally on a par with other models. Regulatory authorities in the UK do not currently adopt model performance thresholds as pass/fail criteria, i.e., there are no formal requirements to show that a model meets certain performance thresholds for that model to be used in practice. Instead, model performance metrics are seen as a potentially useful aspect of a submission to regulators that can be used to help evaluate how well a model performs against other models and measurements. In other jurisdictions, metric thresholds are employed by regulatory agencies in a more direct way. An example being the model evaluation protocol used by the US Pipelines and Hazardous Materials Safety Administration [3, 4].

Model evaluation workshop: Thresholds and toolkit

To share the findings of the CERC project with the dispersion modelling community, the ADMLC funded a workshop on 11 June 2026 at the UKHSA conference centre in Harwell, Oxfordshire. There were around 50 attendees at the event, including dispersion modelling experts from consultancies, model development organisations, research institutes and academia, as well as users or assessors of dispersion model results from industry and government. The main objective of the workshop was to seek the views from attendees on the threshold values and methodology developed by CERC, i.e., to understand if there was consensus from the dispersion modelling community in using the thresholds in practice to evaluate dispersion models. Facilitated breakout group discussions were held in four groups, with roughly a dozen individuals in each group. The people in the groups were asked to consider several questions about the metrics, the methodology and datasets that CERC had used to calculate the thresholds, as well as their thoughts on whether one threshold or a range of criteria should be used in practice. The feedback from these discussions is summarised below.

Considerations in using thresholds criteria for model acceptance

The attendees were broadly in agreement that the thresholds calculated during this study should not be used to judge if a model is simply “acceptable” or not, in a black-and-white pass/fail way. Several reasons were given for this.

Firstly, the attendees thought that dispersion model performance is complex and dependent on multiple factors. There are many reasons why a model may appear to give acceptable or poor agreement when compared to measurement data. These reasons can relate to: the physics of the scenario; the physics that the model has been developed to represent; the quality of the data available for input to the model; and the quantity and quality of measurement data available for model evaluation. For example:

- Experiments that are conducted in wind tunnels under carefully controlled and repeatable conditions have different uncertainties from field-scale experiments involving measurements made at significant distances downwind (in some cases, thousands of kilometres downwind).
- Complex and perhaps poorly quantified model input data (e.g. time-varying source conditions, meteorology or terrain) may impact the model's ability to reproduce measurements.
- The spatial and temporal resolution of measurements, sensor accuracy and issues with sensor saturation and lower limits of detection can affect recorded measurements
- Routine air monitoring data recorded at hourly intervals over long durations have different characteristics to field and wind-tunnel measurement data for discrete releases.

Attendees at the event explained that factors such as these mean that a model would not be expected to perform equally well across a wide range of different applications. They thought that the application of thresholds is quite case specific: the thresholds should depend on the application, the characteristics of the measurement data, and the purpose of the modelling. A one-size-fits-all set of threshold criteria that is derived from a wide range of datasets representing different applications, from air quality to emergency response and regional/global-scale modelling, is probably not appropriate, nor even necessary. Datasets used as input to this method of calculating model evaluation thresholds should instead be selected carefully to be relevant to the intended use of the model. Even then, different atmospheric dispersion modelling applications may need different thresholds, for instance, requiring more or less stringency.

Secondly, there was a concern that the application of strict model acceptance thresholds for regulatory applications could encourage undesirable behaviour, such as model developers artificially fitting their model to the data. The attendees preferred a more informed and graduated scale, where metrics are used to *explain* model performance within an holistic assessment that clearly describes the capabilities and limitations of both the model and the measurement data. In some cases, it may be deemed acceptable for a model's performance to fall outside strict threshold criteria. For example, fast screening models may not agree well with data, but they may still be deemed useful for the purpose of scoping studies, if they generally provide cautious results.

Thirdly, whilst the CERC methodology for deriving the thresholds is objective and repeatable, some attendees voiced concerns about the principle of thresholds being defined from a set of model predictions, although this approach was previously used by Hanna and Chang. Their concerns were that if some of the models used to define the thresholds were particularly poor models, this could bias the results and lead to artificially wide thresholds. The types of models being used to define the thresholds might also not be representative, i.e., if models were being used outside their range of applicability or if a large proportion of the models were of a certain type. Some attendees

thought that, in some cases, the intended use of the model and/or some features of the measurement data could instead be used to define the thresholds. For example, in safety-critical applications, there may be a requirement to use a model of reliably high accuracy under certain conditions (irrespective of the performance of a range of models against relevant data). Also, a model cannot be expected to perform any better than the range of uncertainty in the measurements.

There was greater support for use of thresholds in air quality applications than in other fields. This appeared to relate to the air quality community having more familiarity and experience with thresholds in support of the European Air Quality Directive and through initiatives such as FAIRMODE. Another example of attendees' experience with thresholds was to support the US Federal regulator PHMSA's evaluation of models used for siting Liquefied Natural Gas (LNG) facilities.

General workshop feedback on model evaluation approaches

Workshop attendees saw a number of situations going forwards where the work undertaken by CERC would have tangible benefits. Model developers stated that CERC's work would be useful in helping them to understand if an adjustment to a model improved its overall performance. The study had identified a large number of different metrics, several of which were new to the attendees. The metrics assessed different aspects of a model's performance, e.g., for bias, scatter, correlation and spatial distribution of errors. There were also metrics that can diagnose possible sources of model error, such as the components of mean square error and RANK. Attendees at the meeting noted that when using a new metric for the first time, it is difficult to know what values to expect from their calculation, and what constitutes good model performance. A benefit of CERC's analyses is that the thresholds provide an indication of suitable values, based on a range of datasets and model predictions.

Attendees also saw merit in the continued use of geometric mean (MG) and the geometric variance (VG) to assess model performance, for example using the classical parabola plot. These measures were not investigated in detail in the CERC work, because the calculation of MG and VG involves taking the logarithm of concentrations, which is problematic when some concentrations are zero. Attendees noted that filtering the data or using aggregated measures such as arc-max concentrations and plume widths can overcome these issues for some evaluation studies.

Another suggestion from attendees at the event was for the ADMLC to share datasets, model predictions, and model evaluation outcomes on its website, to enable modellers to validate their models and benchmark their model's performance. This could include outputs from the CERC Model Evaluation Toolkit, which was extended under this ADMLC project to generate additional metrics relevant to the study. Such data could also facilitate a future study where application-specific threshold criteria could be developed. The ADMLC is open to hosting datasets on its website. If anyone is interested in sharing their data, please do not hesitate to contact us at: admlc@ukhsa.gov.uk.

Summary

The CERC report and workshop presentations have improved the community's understanding of model evaluation metrics. The ADMLC is very grateful for the diligent and comprehensive approach CERC took to undertaking their work, and to the useful input provided by Steven Hanna, who collaborated with CERC. The ADMLC would also like to sincerely thank all of the attendees at the workshop for their useful input to the discussions in the breakout groups.

The ADMLC committee members are drawn from organisations that consider a range of different applications of atmospheric dispersion modelling. If represented on a Venn diagram, the members' interests would feature a core region where we share a common purpose and satellite regions where individual organisations focus their efforts. On reflection, the ADMLC presented CERC with a difficult challenge in this project of applying model evaluation metrics to such a wide range of application areas that spanned the topics of interest to the committee members. CERC met the challenge and it is now perhaps a task to be taken forward by each of the ADMLC member organisations, to apply the knowledge gained from this project in our respective domains. We expect the CERC report to be used for many years to come as a valuable reference document on the subject of model evaluation.

The ADMLC Committee, July 2026

References

[1] Chang, J.C. and Hanna, S.R. (2004). Air quality model performance evaluation. Meteorology and Atmospheric Physics, 87, 167–196. https://doi.org/10.1007/s00703-003-0070-7

[2] Hanna, S.R. and Chang, J.C. (2012). Acceptance criteria for urban dispersion model evaluation. Meteorology and Atmospheric Physics, 116(3), 133-146. https://doi.org/10.1007/s00703-011-0177-1

[3] US Code of Federal Regulations 49 CFR 193.2059 "Flammable vapor-gas dispersion protection" https://www.ecfr.gov/current/title-49/subtitle-B/chapter-I/subchapter-D/part-193

[4] Pipelines and Hazardous Materials Safety Administration (PHMSA) guidance on "LNG Exclusion Zones-Radiant Heat Flux & Vapor Dispersion Modeling" https://www.phmsa.dot.gov/pipeline/liquified-natural-gas/lng-exclusion-zones

[5] Ivings, M. J., Lea, C. J., Webber, D. M., Jagger, S. F., & Coldrick, S. (2013). A protocol for the evaluation of LNG vapour dispersion models. Journal of Loss Prevention in the Process Industries, 26(1), 153-163. http://dx.doi.org/10.1016/j.jlp.2012.10.005