19.1 INTRODUCTION

Since about 1975, much research has been devoted to consequence modeling. This has involved the modeling, mathematically as well as physically, of the chemical and physical phenomena associated with major industrial hazards (MIH). Such models are used primarily in risk assessments for safety reports and by safety officers. These models may therefore influence very important decisions, such as the design or authorization of chemical plants. Proper attention should therefore be paid to the quality of these models.

The quality of consequence models, especially dense gas dispersion models, started to be assessed around 1980. From that time, experimental data, both at laboratory scale and at field scale, have been gathered. Among the more famous large-scale experiments are those performed by Lawrence Livermore National Laboratory in the United States (with names like Desert Tortoise, Coyote, and Burro) and those by Shell and UK Health and Safety Executive in Europe (Maplin Sands, Thorney Island). These experiments resulted in a considerable improvement of the available models, reducing the range of variation between the predictions of the different models (McQuaid, 1983).

As modeling capabilities increase and more complex problems are addressed, however, there remains a serious concern related to model quality and the possible limits for areas of application. Since the Thorney Island trials, the European Commission (Directorate General XII, EC DGXII) alone has supported two major series of field tests in parallel with wind tunnel and modeling work. In addition, there are many heavy gas dispersion (HGD) models of varying quality and applicability, ranging from simple box models through more sophisticated shallow layer-type models to fully three-dimensional, computational fluid dynamics (CFD) models. Each type of modeling has its merits and disadvantages, and the abundance of models presents a baffling situation for those having to choose and use models. A structured and unified approach to evaluation and quality enhancement is important in this field.

An initiative was taken by the EC DGXII, which led to the study carried out by Britter (1992). This study provided a framework for managing the quality and evaluation of technical
models. As a result, a Model Evaluation Group (MEG) was set up by the European Commission. The group produced a generic evaluation protocol for consequence models (EC, 1994a).

In 1993, the Model Evaluation Group set up an expert group on heavy gas dispersion, one of the aims of which was to produce an evaluation protocol specific to HGD models. The protocol produced by the group was tested in a small evaluation exercise. This protocol was taken further by the project Scientific Model Evaluation Techniques Applied to Dense Gas Dispersion Models in Complex Situations (SMEDIS) (Daish et al., 1999). About 30 different institutes contributed to this project, with their models supported by the European Commission.

### 19.2 EVALUATION

Several evaluation studies and model intercomparison studies have been done since 1980 (e.g., Mercer, 1988; Hanna et al., 1991; Brighton et al., 1994). Hanna et al. (1993) carried out a seminal validation study on a number of commonly used dense gas dispersion models. This study focused on flat, unobstructed terrain and included only some elements of scientific assessment. These evaluation exercises focused on comparison of calculated data and measured data, using some statistical measure to express the degree of agreement between the different sets of data.

Like Hanna (1991), we want to stress beforehand that a statistical evaluation is only one part of the evaluation of dispersion models. Evaluation of the scientific basis, solution techniques, and the model description is at least equally important in assessing the validity or, more generally, the usefulness of a model.

The purpose of model evaluation is twofold. For use in an operational environment, it can assist interested parties in choosing a correct model and interpreting output results. More precisely, model evaluation should do the following:

- It should produce measures of model quality that may be communicated to users and other interested parties. The evaluator (which is the editor of the evaluation document) ensures that the results of the model evaluation are accessible to non-expert users. The document provides the interested party with the information on the bounds of applicability.
- It should allow a referee to consider the applicability of a model to specific situations and to assess the weight that should be given to the model results.

The second purpose of model evaluation is to encourage appropriate model improvements in a cost-effective way. An evaluation can:

- Encourage the management of model quality
- Reduce distortions in output results by the use of different models
- Identify possible areas of improvement in a model
- Help to identify shortcomings in data sets requiring further experimental effort

Model quality needs to be evaluated and communicated to interested parties in a structured way in order to ensure acceptance and usefulness. The MEG advises a structure consisting of three elements: assessment, verification, and validation (EC, 1994b). In this chapter, we adopt the terminology as suggested by the MEG protocol, i.e., model evaluation includes the whole review of any model with respect to proper scientific formulation (assessment), correct coding (verification), and comparison with experimental data (validation). These stages are defined in Table 19.1. We are aware that other terms are sometimes used and that there is some philosophical criticism of the use of the term validation.
TABLE 19.1 Definition of Terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model evaluation</td>
<td>This includes examination of a model according to a set of well-defined rules embodied in the evaluation protocol. Although the scientific assessment of the model is the central activity, “evaluation” is used to mean the entire range of activities before, during, and after the scientific assessment.</td>
</tr>
<tr>
<td>Validation</td>
<td>This is the process of comparing the predictions of a model, which has been run to simulate a given event, with the observations made in connection with the same event.</td>
</tr>
<tr>
<td>Verification</td>
<td>This is the process of comparing the implementation of a model with its mathematical basis. Most commonly, this refers to checking that a computer implementation of a model (computer software) is an accurate representation of the algorithms in the model.</td>
</tr>
</tbody>
</table>

The assessment is itself composed of two parts. The scientific assessment identifies the mathematical and physical algorithms of the model. The user-oriented assessment concerns operational aspects such as costs and ease of use.

Verification provides evidence that the model has been checked on a correct coding of algorithms, databases, and interfaces.

Validation is described in the MEG documents as “getting the model to predict experimental results.” The validation supplies the evidence that the output results of the model compare well with experimental data. The usefulness and accessibility of the experimental data are quality aspects of the comparison.

Typical stakeholders in model evaluation are the model developers, safety officers in industry and consultants applying models during preparation of safety reports, and authorities facing approval of these safety reports. The preferable approach is that either a model developer or a user is able to undertake the evaluation. An independent agent with appropriate expertise can also undertake the evaluation. Because the question of impartiality may be raised, the relation between the evaluator and the model should be documented. In each case, however, the evaluation report should be fully verifiable (e.g., by means of an audit).

Prior to any evaluation exercise, the context, scope, and aim of the exercise should be made explicit. It is important to define the area of application for the evaluation in relation to the originally intended use of the model. It is also important to consider a priori the use of the evaluation results.

These aspects and the other procedures to be followed during the exercise will be laid down beforehand in an evaluation protocol. The next two sections will provide more details on the scientific assessment and the validation stages. An outline of a general evaluation protocol for dense gas dispersion models will be presented and some previous evaluation studies will be discussed.

19.3 SCIENTIFIC ASSESSMENT

Until now, most evaluation studies and model intercomparisons relied on comparison of the outputs of models with observations or other model results. However, there is a limited amount of observations of sufficient quality compared to the complexity of the phenomena, not to mention the inherent variability of atmospheric dispersion, which makes it difficult to measure. Therefore, such validation studies alone do not provide sufficient proof of a model’s quality and its capabilities to address problems outside the range of the validation data sets. A scientific assessment of the model is an important means to obtain information about the capabilities, reliability, and quality of a model. The scientific assessment seeks:
1. To clearly identify the scientific basis of the model and how this scientific basis is implemented

2. To provide an overview of the capabilities and limitations of the model, based on the chemical/physical aspects (“features”) that are explicitly addressed in the model

3. To provide a judgment on the adequacy of the scientific basis and its implementation in relation to the present state of the art both for the model as a whole and for the individual features

Items 1 and 2 can be achieved by writing a comprehensive description of the physical, chemical, and mathematical basis of the model in a unified format or protocol that systematically addresses the different features relevant to heavy gas dispersion. A unified format, in the form of a template to be filled for each model, will also enable an easy comparison between different models. The effort required to produce such a description depends on the quality and completeness of the available documentation of the model. Hopefully, future model developers will follow the format of the protocols that have been developed by now. In any case, it is to the advantage of a model if the original documentation is as detailed and comprehensive as possible.

Performing the review as intended under item 3 is a more problematic task. Competence to judge scientific quality requires a sound expert knowledge of dense gas dispersion as a specific scientific domain linked to fluid dynamics, thermodynamics, and meteorology. Generally accepted scientific principles are based on verification and falsification of theories and assumptions. However, nowadays dense gas dispersion models are complex systems that cannot be simply verified or falsified. Scientific assessment requires that the following aspects be considered carefully:

- The traditional method of scientific assessment is the peer review. It is recommended that the evaluator set up a peer review panel (where the evaluator can be a member, but not necessarily) to review the documentation produced during the assessment. It should be recognized, however, that peer reviews are subjective and tend to be conservative, i.e., innovative solutions may be unjustly rejected.
- Detailed descriptions and validations of single parts of a model in its original documentation, e.g., using analytical solutions or specific well-controlled experiments, may be in favor of the model, showing the thoroughness of the developer in developing the model.
- For the same reason, the amount and quality of the literature to which the model developer refers and the developer’s argumentation for certain approaches as compared to alternative methods may be in favor of the model.
- The number and quality of publications based on the model and citations to these publications may give an indication of the model’s quality. With respect to the quality of a publication, peer-reviewed articles in international journals are normally ranked highest. It provides a means to include the referees’ judgment in the review. Whether or not a model is published in a refereed article therefore makes a differences, but the correlation between scientific quality and the number of international articles is low in general. One should also consider that a citation does not necessarily mean an approval of the cited material (Christiansen and Hansen, 1993).

19.3.1 Review of Capabilities and Physical–Chemical Aspects

The structured description of the model provided by the scientific assessment is especially useful for addressing the capabilities of the model. General capabilities relate to the source properties, the topography, and atmospheric conditions:
• Duration, size, and momentum of the source
• Contents of the cloud, thermodynamic properties and chemical composition of the released material, aerosol behavior, chemical reactions, deposition, and heat transfer
• Domain geometry (slopes, buildings, and similar obstacles)
• Atmospheric conditions (stability, humidity, wind speed, and turbulence)
• Ground conditions (surface roughness and heat fluxes)
• Minimum and maximum distances from the source

Aspects to be included in the scientific assessment depend on the type of model. Models can be distinguished on the basis of the spatial dimensions of the independent variables:

• 1D (downwind distance or travel distance/time is the independent variable, lateral similarity profiles of concentration are often included)
• 2D (either two horizontal dimensions—typical for shallow layer models—or downwind distance or travel time and plume height—typical for plume trajectory models—are independent variables)
• 3D (explicit calculation in three directions)

Models can also be distinguished according to the solution technique:

• Phenomenological models and screening tools, in which the dispersion behavior is described by a series of nomograms
• Integral models (e.g., box models for instantaneous releases and steady plume models for continuous releases)
• Shallow-layer models
• Computational fluid dynamics CFD models

The following are the categories of physical and chemical aspects to be reviewed for all classes.

• General:
  • Parameterization of processes and possible discontinuities introduced when different expressions are used
  • Limiting or asymptotic behavior, e.g., dense to passive; short finite release to instantaneous release

• Fluid dynamics and diffusion processes:
  • Representation or calculation of atmospheric flow, including its turbulence and diffusion characteristics
  • Turbulence closure used, if any
  • Formulation of interaction of atmospheric flow with terrain and/or obstacles, including friction and modification of the flow and turbulence
  • Coupling with other dispersion regimes, especially neutral and positively buoyant ones
  • Representing source geometries by approximating them with available source-term simplifications in the model
  • Modeling approach: Lagrangian/Eulerian treatment, deterministic/stochastic treatment
  • Concentration fluctuations, plume meander
  • Averaging considerations—averaging assumed in model, averaging time for output
CHAPTER NINETEEN

TABLE 19.2 Phenomenological Models and Screening Tools

In general, the models provide an empirical relationship between the concentration decay of the gas cloud and the downwind distance from the source, in terms of dimensionless quantities. The model type is characterized by the choice of included parameters. The most important parameters are:

- The choice of scaling for dimensionless quantities
- Gravity constant
- Gas density relative to ambient air
- The released volume (instantaneous) or release rate (continuous)
- Ambient wind velocity

**Thermodynamic processes:**
- Formulation of heat transfer (sensible and latent) from various types of surface to the cloud
- Formulation of treatment of contaminant aerosol, including possible interaction with ambient water (vapor and liquid)
- Correlations for thermodynamic properties, e.g., density, vapor pressure

**Chemical and other processes:**
- Formulation of chemical reactions, whether spontaneous or with the environment (e.g., water vapor in atmosphere reacting with HF)
- Correlations for chemical properties
- Radioactive decay

**Initial/boundary conditions:**
- Comment on characterization of source, including any link between dispersion model source and different types of primary source (jet, leakage, etc.)
- Coupling with source term model

Tables 19.2 to 19.7 provide checklists of the physical aspects that need to be addressed for the various models.

19.3.2 Outcome of the Scientific Assessment

The scientific assessment leads to a structured presentation of the scientific basis of the model. This structured presentation allows a systematic comparison of the capabilities of different models based on the features and aspects that are explicitly taken care of in the model.

The outcome of the scientific assessment preferably also includes an expert judgment of the scientific basis of the model and its implementation. Such an assessment will always be subjective because objective indicators for scientific quality do not exist. Therefore, the model developer/provider should always be given the opportunity to comment on the assessment as part of the evaluation report.

19.4 VALIDATION

There is much more experience with validation than with scientific assessment. This validation can either be qualitative, for example by means of scatterplots of observed versus modeled concentration, or more quantitative, usually by means of statistical measures of the difference between the two.

**TABLE 19.2 Phenomenological Models and Screening Tools**

In general, the models provide an empirical relationship between the concentration decay of the gas cloud and the downwind distance from the source, in terms of dimensionless quantities. The model type is characterized by the choice of included parameters. The most important parameters are:

- The choice of scaling for dimensionless quantities
- Gravity constant
- Gas density relative to ambient air
- The released volume (instantaneous) or release rate (continuous)
- Ambient wind velocity
TABLE 19.3  Box Models

Box models are used to describe instantaneous heavy gas releases (puffs). Physical aspects of interest for a box model are:

• Entrainment of air into the cloud.
• The expressions for entrainment into the edges and/or the top of the cloud are often different in different models and should be reviewed.
• Heat transfer between the substrate and the cloud. The transfer can be a sensible heat flux or a latent (evaporative) heat flux. In the latter case, it can be the evaporation of aerosols in the gas or the condensation of ambient water vapor.
• Chemical reactions.
• The vertical wind profile and atmospheric stability.
• The frontal spreading velocity.
• The form and the length scale used for both the horizontal and vertical concentration distributions.

In some cases the box model approach has been extended for the simulation of a continuous or time-varying release. The total release is then divided into a number of puffs, each of which is considered as a separate release. The model evaluator should then discuss the modeling of gravitational slumping and mixing in the along-wind direction.

TABLE 19.4  Steady Plume Models

Steady plume models are developed in the same way as box models, with the variables depending on the downwind distance instead of time. Physical aspects of interest are the same items as in Table 19.3 for box models. The following are some other aspects for steady plume models:

• Time-varying releases are simulated by a superposition of steady plumes, which are formed at subsequent time intervals. The model evaluator should then discuss the modeling of gravitational slumping and mixing in the along-wind direction.
• The steady plume models cannot describe instantaneous releases because the initial slumping phase is not incorporated. This can be remedied by adding a front-end box model for the gravity-dominated stage.
• The transition to passive dispersion.
• The similarity profiles of the heavy gas concentration in the plume and relation to the algorithm of air entrainment.

Although readily available procedures for validation studies do not yet exist, when comparing model results with experimental data, three major problems occur:

1. The level of spatial and temporal detail one wants to retain in the comparison
2. The selection of experimental data
3. The statistical measures used to quantify the result of the comparison

19.4.1 Spatial and Temporal Detail

The highest level of detail occurs if one tries to compare model results and experimental data paired in time and space, i.e., each point value from a series \( C_{\text{observed}}(t, x, y, z) \) is compared with the corresponding point value from the series \( C_{\text{model}}(t, x, y, z) \). Due to the
TABLE 19.5  Integral Plume Models

The integral plume theory is intended for application on lofted plumes. Physical aspects of interest are:

• Entrainment of air into the cloud. Points to check are (1) the presence or absence of an entrainment term involving atmospheric turbulence and (2) the presence or absence of an entrainment term due to plume turbulence far away from the source.

• The inclusion of a drag force in the conservation equations of momentum.

• The presence of a zone of flow establishment.

• The shape of the plume cross section. The models employ either a circular or elliptical cross section.

• The profile of the crosswind variables around the centerline of the plume

• Chemical reactions

• The vertical wind profile and atmospheric stability.

inherent variability of dispersion in the atmosphere and therefore the fundamental impossibility of specifying initial and boundary conditions well enough to be able to reproduce $C_{\text{observed}}(t, x, y, z)$ completely, ensemble averages have to be constructed in the form of some processing of the data, both model data and experimental data.

Due to limited resolution in experiments, some averaging over time and volume is involved. Normally, some time averaging is required also to average out turbulence and wind-direction fluctuations that cannot be resolved experimentally and in the boundary conditions for the model.

Another simplification that is often applied is to compare some derived characteristics as a function of distance from the source, mainly the maximum concentration (often called maximum arc-wise concentrations) and the crosswind width of the plume or cloud. For instantaneous releases, the dose (integral of concentration over time) can be used instead of concentration. This procedure can be very effective because the major concern related to heavy gas dispersion is the assessment of hazard distances and the area affected by the released material. Duijm et al. (1996) conclude that for a proper conclusion of the overall

TABLE 19.6  Shallow-Layer Models

The shallow-water method is based on vertical integration of the cloud properties, reducing the problem to a two-dimensional set of conservation equations (mass, species, downwind and crosswind momentum, and energy). Some models consider only downwind distance as the independent variable. Physical aspects of interest are (see also Table 19.3 for box models):

• Entrainment of air into the cloud.

• The frontal spreading velocity.

• The transition from plume to puff mode. Close to the source the release can resemble a plume, but the model switches to transient puff mode when the steady-state period is over.

• Heat transfer between the substrate and the cloud.

• Chemical reactions.

• The vertical wind profile and atmospheric stability.

• Shear stress. The conservation equations contain downwind and crosswind friction terms.

• Vertical concentration profiles and the definition of cloud height.
TABLE 19.7  CFD Models

<table>
<thead>
<tr>
<th>Physical aspects of interest for CFD models are:</th>
</tr>
</thead>
<tbody>
<tr>
<td>• The turbulence closure model. Different levels of turbulence closure are used to obtain the eddy diffusion coefficients. The turbulence models differ in the assumptions on isotropy of turbulence.</td>
</tr>
<tr>
<td>• The simplifications in the conservation equations. Examples of different approximations are: hydrostatic, boussinesq, anelastic.</td>
</tr>
<tr>
<td>• The type of domain mesh: structured or unstructured grid.</td>
</tr>
<tr>
<td>• The discretization method of the equations. Finite-difference methods, finite-element, and finite-volume methods are all used.</td>
</tr>
<tr>
<td>• Terrain/obstacle effects refer to obstructions in the flow field and to variations in the terrain height.</td>
</tr>
<tr>
<td>• Meteorological fields used as initial and boundary conditions.</td>
</tr>
<tr>
<td>• Momentum, heat, and mass (e.g., deposition) transfer between the cloud and its surroundings.</td>
</tr>
</tbody>
</table>

behavior of the model, it is necessary that comparison of maximum arc-wise concentration or dose be combined with comparison of the plume width or cloud width and length, respectively.

Using arc-wise concentrations and plume width also allows the random effect of wind-direction variations to be circumvented by using the concept of moving frame averaging of the experimental data (if sufficient crosswind sensors are available) (see Nielsen and Ott, 1996).

For “simple” situations where plume direction is dominated by wind direction only, using arc-wise maximum concentrations together with plume width is the recommended procedure. However, for other situations, such as dispersion close to buildings or in complex terrain, part of the model performance depends on the correct prediction of the plume path, or there may not even be a unique plume path. In those situations, comparisons of concentration or dose paired in space should be included. In those cases, the experimental data points (sensor positions) included in the comparison should be restricted to points that give significant information about the location and behavior of the plume or cloud, i.e., the number of zero-level measurements should be restricted, though zero-level measurements contain valuable information.

19.4.2 Data Set Selection

The selection of the data set depends primarily on the type of model and the aspect of the model that one wants to test, e.g., dispersion over flat terrain (a considerable volume of data exists for this situation), dispersion close to buildings, or dispersion in the presence of heat transfer.

The data set should preferably cover a reasonably wide spatial range in a structured way. To allow comparison with arc-wise maximum concentrations, the sensor arrangement should include at least four crosswind or arc-wise sensors with less than one standard deviation of plume width between them, although more than six sensors per arc are recommended. Preferably there is at least one position at each arc where vertical concentration distributions are available. For field data, frequency responses of sensors between 0.1 and 1 Hz are fine for general use (some averaging will often be performed afterwards), unless concentration fluctuations are part of the validation. The relevant time scale for sensor response is the time of flight $u/x$ at a downwind distance, $x$.

Attention should be paid to the origin and quality of the data and, if secondary sources are used, one should be aware of any data processing that has been applied, such as time averaging and rejection of certain data series. The REDIPHEM database (Nielsen and Ott,
19.10 CHAPTER NINETEEN

1995) contains one-second block-averaged data, but otherwise it keeps very close to the original data sets. In the modelers data archive by Hanna et al. (1991) the information is much simplified and therefore convenient to handle. It consists of tables of the highest average concentrations observed at various distances and plume widths based on average concentrations. For a discussion of quality aspects of data sets and documentation, see Nielsen and Ott (1995). Other information on available databases can be obtained from Mercer et al. (1998) and Carissimo et al. (2001).

An essential quality factor is the reliability of the information on boundary conditions, such as wind speed, stability, and local surface roughness. The reliability of the source term and release rate information is extremely important. Doubts on the real value of release rates have in fact reduced the potential of some large-scale experiments.

Another issue related to data selection is whether to include laboratory data, especially results from wind-tunnel experiments. The advantages of wind-tunnel results are that boundary conditions are well controlled, large volumes of good quality data are available, and specific aspects, especially effects of obstacles such as buildings, can be studied in detail. On the other hand, meteorological conditions, such as wind meandering, are not fully “real” (this may also be an advantage), there may be scaling effects related to small values of the Reynolds number, and heat transfer, droplet evaporation, and chemical processes cannot normally be reproduced.

19.4.3 Statistical Parameters

Analyzing the performance of a dispersion model by comparing the model output with observed values should result in the establishment of some performance measure. Often such a performance measure consists of one or more statistical parameters, but it can also be a written statement or a graphical presentation. Here we restrict ourselves to statistical measures, keeping in mind that they should be applicable to whole data sets as well as to specific subsets of data in order to examine trends of the performance with some external parameters (e.g., atmospheric stability or downwind distance).

Specifying “performance” of a model in an objective way is not a straightforward process. It can have several aspects. We specify the following requirements;

• A performance measure should indicate whether the model in general under- or overpredicts.

• A performance measure should indicate the level of scatter, i.e., the random deviation from the average under- or overprediction.

If models are evaluated over the full downwind range, i.e., over a range of concentrations varying by some orders of magnitude, the performance measure should weight all pairs of observation and prediction equally, independent of the absolute value of the concentrations. This is the preferred method if a validation is based on arc-wise concentrations and plume width. If the validation is based on spatially paired data points, the disadvantage of this method may be that mismatch between model and observations at the edge of the plume may dominate over relatively good agreement at the center of the plume. Performance measures that consider the absolute difference between model calculation and observation can be used if the range of maximum concentrations in the database does not change too much, e.g., when one is interested in the concentrations close to a single building. A discussion of different performance measures can be found in Duijm et al. (1996).

A pair of performance measures that considers the absolute difference between model and observations are the fractional bias and the Normalized mean square error (see Hanna et al., 1991), which are only “normalized” by the data set-averaged concentration. Averaging over the data set is indicated by the operator < >:
Fractional bias: \( FB = 2 \frac{\langle C_p \rangle - \langle C_o \rangle}{\langle C_p \rangle + \langle C_o \rangle} \)

Normalized mean square error: \( NMSE = \frac{\langle (C_p - C_o)^2 \rangle}{\langle C_p \rangle \langle C_o \rangle} \)

In view of the requirement of equal weight irrespective of the absolute value of the concentration, Duijm et al. (1996) introduced the mean relative bias, together with the mean relative square error:

Mean relative bias: \( MRB = \left( 2 \frac{C_p - C_o}{C_p + C_o} \right) \)

Mean relative square error: \( MRSE = \left( 4 \left( \frac{C_p - C_o}{C_p + C_o} \right)^2 \right) \)

The MRB ranges (similarly to the FB) from \(-2\) to \(2\), with an optimum value of 0. A negative MRB indicates underprediction and a positive MRB indicates overprediction.

The above measures are based on the difference between observation and prediction. Another approach is to consider the ratio of the observation and the prediction \( C_o/C_p \). This puts inherently equal weight on the pairs irrespective of the absolute values.

The simplest set of performance measures using ratios is based on sign tests. To those belong the fraction of overprediction (the fraction of pairs for which \( C_p > C_o \)) and the factor of \( n \) statistics (the fraction of pairs for which the ratio is between \( n \) and \( 1/n \)):

Fraction of overprediction: \( FOEX = \frac{N_{C_p>C_o}}{N} - 0.5 \)

Factor of \( n \): \( FAn = \frac{N_{1/n<C_p/C_o<n}}{N} \)

The FOEX ranges from \(-50\%\) to \(+50\\%\), with an optimum value of 0. The FAn ranges from an optimum value of \(100\%\) to 0. The factor of 2 (FA2) is most often referred to in dense gas evaluation studies (Hanna et al., 1991). The sign tests have the advantage of being distribution-free.

One of the evaluation systems developed by Hanna et al. (1991) is based on the geometric mean bias (MG) and the geometric mean variance (VG). By first taking the logarithm, extreme ratios have a less than proportional weight in the final result:

Geometric mean bias: \( MG = \exp(\langle \ln(C_o/C_p) \rangle) \)

Geometric mean variance: \( VG = \exp(\langle (\ln(C_o/C_p))^2 \rangle) \)

The MG varies between 0 and \(\infty\) and the optimum value is 1. Values smaller than 1 indicate overprediction by the model and values larger than 1 indicate underprediction. In order to obtain symmetry in the final results, we prefer \( \ln(MG) \), which ranges from \(-\infty\) to \(\infty\) with an optimum value of 0. In this number, a same amount of under- or overprediction leads to the same absolute values (but with opposite sign). VG is the variance counterpart of MG.

The advantage of the MRB over the MG is that the MRB accepts if one of \( C_o \) or \( C_p \) has a zero value. Data pairs where either \( C_p \) or \( C_o \) is zero contain valuable information, although instrument properties should be considered when interpreting low-level measured concentra-
tions. The MRB is discriminative for $C_p/C_o$ ratios between 0.1 and 10 but little outside this range.

Duijm et al. (1996) compared different performance measures. Based on this study, the MRB, MRSE, and FA2 parameters are recommended, though the MG and VG can replace the MRB and MRSE. These parameters lead to the most consistent ranking of model performance. The FA2 is an easy-to-understand measure and takes into account the effects of bias and variance in a straightforward manner.

The total volume of experimental data is still small compared to the variation in conditions, and it appears that the numeric results of the performance measures depend on the data sets selected. Therefore, models and model performance numbers should always be compared using exactly the same data sets. The value of the performance measures have little absolute meaning, i.e., one should never rank models on performance measures obtained from different evaluation exercises.

In some cases it can be useful to group data sets to improve the statistical significance of the results. For example, in SMEDIS, data sets have been grouped by category of complex effects (aerosols, obstacle, terrain) (Carissimo et al., 2001).

A statistical model evaluation exercise cannot be based on assessing performance numbers alone. One should try to explain why certain models respond in a certain way. Graphical displays can be helpful to explore the characteristics of models and their behavior for certain data sets and provide additional information to the performance numbers.

## 19.5 EVALUATION PROTOCOL

The protocol described in this section is based on (though not identical to) the heavy gas dispersion model evaluation protocol developed by the Heavy Gas Dispersion Expert Group (Mercer et al., 1998). This protocol provides guidance on how a dispersion model should be evaluated and the steps that must be taken. It follows the guidelines of the Model Evaluation Group (EC, 1994b) for structuring an evaluation protocol and extends these to heavy gas dispersion modeling. A protocol can be shaped to contain more specific considerations and actions. An example of such a protocol was developed during the SMEDIS project (Daish et al., 1999).

The protocol divides the evaluation procedure into six steps:

1. Model description
2. Database description
3. Scientific assessment
4. User oriented assessment
5. Verification
6. Validation

These steps are described in each section of the protocol together with guidance and practical information on the steps. The evaluation report should preferably follow this protocol. In the introduction to the evaluation report, the relation between the evaluator and the model should be explained, as well as the aim, context, and framework of the evaluation.

### 19.5.1 Model Description

The model description provides a brief identification of the model to the interested party. The interested party is either a (potential) user or a referee. The identification text contains the following information.
The name, version number, and release date: It is essential that the dispersion model be clearly identified by name and version number.

The area of application: The applicability of a dispersion model for a particular risk analysis problem depends on the actual conditions in the near and far field.

The name, address, telephone and fax number of the originating organization.

The source and costs of the model.

The model type: Heavy gas dispersion models can be divided into different classes according to the number of dimensions of the independent variables (1D to 3D) and/or the scientific basis and solution technique (e.g., integral model).

Hardware requirements: Information must be provided on the following hardware:

- The processor (personal computer, workstation, minicomputer or mainframe)
- Memory requirements (RAM)
- Storage device requirements (disc space)
- Other devices (visual display units, graphic cards, printers, plotters, and mice)

Software requirements: The software can impose requirements on:

- The operating system, identified with name and version number
- Graphic packages
- Drivers

Quality assurance: Reference should be made to the guidelines or standards used during development of the model and software, such as the EC guidelines for model development or international quality standards.

A list of references to relevant publications: The list of references should refer to all relevant model documentation: user manuals, tutorials, and technical reference manuals. The description should also provide the interested party with relevant references of publications in the open literature. Heredity, i.e., the relation between the model and other and previous models, is of particular importance in the area of heavy gas dispersion modeling. The lists of references should contain at least the scientific source of the algorithms in the model.

19.5.2 Database Description

The database description identifies the data to be included in the validation and describes the most important aspects. The database description is composed of two forms. The identification form contains the following information for each data set.

Identification Form

Data Type. The text should indicate the name or identifier of each data set and the date of data collection. The data sets may originate from several sources. For each source, the type should be indicated. It can be:

- Analytic results
- Results from existing (more sophisticated) models
- Laboratory experiments
- Large-scale field experiments
- Accident reports
**Data Ownership and Accessibility.** The intention is a text that clearly states where the data resides and in what format and who owns the data.

**Origin of the Data.** Because much useful data already exists in databases, the database description should contain, wherever possible, the following two statements for each data set:

1. A statement that allows the data to be traced back to a primary source. The source of the original data should be clear
2. A statement that specifies the possible loss of information between the primary source and the presented data set. The statement should mention any processing of the data that has been carried out: deletion of data points, truncation of time series, etc.

**Quality Assurance.** The intention is to describe what quality assurance (if any) has been performed on the data and any uncertainties about the data. Examples of topics that are addressed here are:

- The calibration of the detectors
- The correction of data for detector errors
- Error bars or uncertainty values for the measured quantities
- An estimate of the inherent variability in the measured variables

**Second Form**

A second form is added to the identification form in order to state clearly any processing of the data that has been carried out and the reasons for using the data set in the validation. This text contains the following information.

**The Appropriateness of the Data.** A statement must be given that explains why the data set is acceptable for use in the model evaluation. Any properties of the data that limit their use in the evaluation process must be stated. Properties of the data that make it particularly useful in the evaluation process should also be stated.

**The Features and Parameters Covered by the Data Set.** A statement is required to specify the ranges of input parameters of the model that are covered by the database. This information could be in a raw variable form or in a dimensionless form. A stepwise approach is suggested. The first step is to identify the component features in the model. A feature defines a scientific area, such as:

- Physics of the atmosphere
- Heavy gas dynamics
- Thermodynamics
- Chemical reactions

The second step is to determine the important parameters for each feature.

### 19.5.3 Scientific Assessment

The scope of the scientific assessment must be clearly stated. This statement provides the user with information on any of the model’s aspects that have been assessed. A motivation for the assessment should be included. In some cases it may be desirable to evaluate isolated aspects of a large model, in which case the method used to isolate the aspects should also be discussed.

The scientific assessment is made up of the following:
A Comprehensive Description of the Model. The scientific assessment by the model evaluator supplements the information in the model description, which is usually based on the documentation from the model developer. The assessment should:

- Be limited to documenting the dominant underlying scientific basis. Peripheral parts of the model are split off prior to this activity.
- Review the algorithms for the dominant process. This can be the complete dispersion algorithm or a particular aspect.
- State explicitly the values of any empirical constants in the model. The constants may have been tuned at the development stage to give the best possible agreement with one or more data sets. If so, the model evaluator should identify the data sets.

An Assessment of the Scientific Content. The scientific content of the model should be assessed. Based on the description of the model, the evaluator is required to provide a judgment of the adequacy of the underlying scientific basis and its use in the model, for example compared to the current state of the art. The evaluator should have knowledge of and access to additional information in scientific review papers and other publications in the literature.

Limits of Applicability. The limits of applicability of the model should be discussed. This is based on an assessment of the features that are explicitly included in the model and the adequacy of the modeling of these features. These features can also include integrated source term models.

Limitations and Advantages of the Model. In discussing the limitations and advantages of the model, it is sometimes necessary to document particular positive and negative aspects of the model. For example, models may range from a screening tool to a research tool, or a model can be designed to be always on the conservative side. If this is the case, then the statements have to be preceded by clear and sensible arguments.

Any Special Features. The formulation should be described for any special features, e.g., the capability to deal with obstacles and/or complex terrain.

Possible Improvements. Suggestions for possible improvements in the model are best included in the scientific assessment.

19.5.4 User-Oriented Assessment

A “user” could be a person in industry (or a consultant) responsible for carrying out a safety analysis and/or preparing a safety report, or a person in a regulatory authority responsible for assessing such safety reports and/or carrying out risk assessments. It is important to state the aims and the scope of the user-oriented assessment and discuss any limitations of the model. The statement provides the user with information on the model’s aspects that are assessed. A motivation for the assessment should be included. Again, the evaluator should state the type of application that is being made—e.g., as a screening tool or, at the other extreme, as a research tool.

The user-oriented assessment describes how easy the model is to use on aspects such as:

The Associated Documentation. The MEG documents suggest that the user-oriented assessment should cover the aspects “fitness-for-purpose” and “ease-of-use” (EC, 1994a, b). Fitness-for-purpose refers to the degree to which the model is able to provide the necessary
results for solving a specific problem. The documentation associated with the model should discuss fitness-for-purpose in a way that is easily accessible to the user, i.e., the educational level of the documentation has to be assessed by the model evaluator. The documentation should supply example calculations to illustrate the model use and list addresses of the developers, expert groups, or user groups for additional user support.

**Installation Procedures.** The documentation on installation requirements and procedures should be complete and clear and preferably give step-by-step guidance.

**A Description of the User Interface.** Ease-of-use is also a requirement of the user interface. The interface can be interactive (real time), either text based or with the help of a GUI, or batch oriented.

**Guidance in Selecting Model Options.** An important aspect for evaluation is the degree to which the model’s user manual or the help facility (in case of a computer implementation) gives guidance in selecting model options. Other means to assist the user are computer-oriented training facilities and short courses.

**Guidance in Preparing Input Data.** The user manual of the model or the help facility (in case of a computer implementation) should give guidance in the preparation of input data. Aspects that can be addressed are:

- The degree to which the input processor checks on typing errors and allows the user to make corrections.
- Initiation of variables with default values.
- The presence of a log file allows the user to check the input data. For batch-oriented as well as for interactive calculations, it is advised that the log file can also be used as an input file.
- A facility for archiving of input/output data helps the user to store results in a structured way.

The possibility of coupling the model to source term models or the integration of source term models (such as jet models and pool evaporation) should also be addressed here.

**Checks to Verify Whether the Model Is Used beyond Its Scope.** The input processor and the core (dispersion) program should perform checks at two levels: (1) on the validity and range of the inputs and (2) on the obtained values of variable parameters in a program run and should give warnings to the user and eventually stop the calculation.

**Clarity and Flexibility of Output Results.** Complete output results are preferably stored in a file with only a summary displayed on screen. In heavy gas dispersion modeling, the applicability is commonly the description of hazard effects (toxicity, flammability) for use in risk analysis.

Dispersion model output results relevant for continuous releases and clouds of simple plume-like shape are:

- \( C(x) \): the profile of concentration versus distance on the cloud centerline axis
- \( Y(x, c) \): the cross-wind distance to a given concentration level \( c \) as a function of along-wind distance \( x \)
- \( M \): the mass of gas between the lower and upper flammability limits. The quantity is needed for estimating the explosive contents of the cloud.
For more general cloud shapes, the model can supply \( C(x, y, z) \), the concentration distribution in three dimensions. This form of output of the results retains the entire information in the calculation but require the user to perform some postprocessing to obtain \( C(x) \), \( Y(x, c) \), and \( M \).

In time-varying or instantaneous releases, the quantities are all functions of the time \( t \). Two more relevant quantities for time-dependent releases and simple cloud shapes are:

- \( D(x) \): the (toxic) dose versus distance on the cloud centerline axis. The (toxic) dose depends of course on the exposure time, \( t_{\text{exp}} \), being the time between cloud arrival and evacuation. The dose is defined as the integral of \( C \) over the time interval between cloud arrival and evacuation, and the toxic dose is the integral of \( C^n \) over the time interval between cloud arrival and evacuation. The real number \( n \) depends on the toxic chemical substance.
- \( Y(x, d) \): the cross-wind distance to a given dose level \((g)\) as a function of along-wind distance \((x)\).

For more general cloud shape, for example if the cloud is split by buildings, the model can supply \( D(x, y, z) \), the toxic dose in three dimensions.

Relevant concentration levels for toxic levels can be very low, down to 1 ppm, so the output results of the dispersion model should cover the far field.

The transient output can be reduced by taking, for the time period of interest, the maximum occurring values of \( C(x, t) \), \( Y(x, c, t) \), and \( M(t) \)—next to the (toxic) dose \( D(x) \).

Facilities are needed for the postprocessing of output results in order to produce lists, graphs, contour maps of concentration distributions, etc. The facility can either be a dedicated postprocessor or commercially available programs (spreadsheets, databases, etc.).

**Unambiguous and Understandable Error Messages.** Error messages should be self-explanatory or refer to the user manual.

**Computational Costs.** The aspects of computational costs (in terms of running times) can be illustrated in the documentation with benchmark tests for various scenarios. If the program operates in interactive mode, then the response times have to be short.

**Possible Improvements.** Suggestions for possible improvements in the ease-of-use category are best located in the user-oriented assessment.

### 19.5.5 Verification

An assessor has to ensure that the computer code is producing output in accordance with the model specifications. It is emphasized that the undertaking of model verification is an extremely tedious task. In the case of a computer implementation, the task is to prove in a step-by-step notation that the code does what it is supposed to do. The common practice is that developers take a less rigorous approach and perform verification by individually testing sections of the computer code (e.g., subroutines). In addition, developers should illustrate the model implementation by putting comment lines within the code. In general, the model evaluator will appeal to the developer to provide information on the extent the code has been verified.

- It is often possible to carry out checks for internal consistency, e.g., mass and/or mass flux balances.
- It may also be possible to run the code for “simple” scenarios for which an analytic solution may be available.
Also, for such simple scenarios, the results of the code could be compared with the results of any workbook procedures, such as the Britter-McQuaid Workbook (Britter and McQuaid, 1988).

The behavior of the model in certain limiting conditions could also be considered. For example, for a passive release, does the model give results that are consistent with a “simple” passive dispersion model?

Automatic tools can be used in some cases to check the correct types of variables and the correct branching of conditional tests.

19.5.6 Validation

The aims and intent of the model validation should be stated. The text should list the model parameters that are to be tested. The validation is of particular importance because it can provide conclusive evidence for model selection to interested parties. A validation starts with selecting a set of various data sets of known quality and suitable for that purpose (e.g., sufficient number of measurements). At the start one also selects a set of statistical measures defined for the comparison of the model output (such as the calculated peak concentration \( C_p \) of heavy gas at a location) and the observed values (such as the measured peak concentration \( C_o \)).

In the following text, the tasks needed to ensure a satisfactory validation are explained.

**The Database Selection.** An important item to be addressed in the model evaluation is the components from the database that will be selected for the validation procedure. The item includes statements on:

- The adequacy and appropriateness of the database scenarios for the intended validation
- Treatment of source terms and of the meteorology
- The inclusion of laboratory and field experiments within a single validation
- The scaling aspects of laboratory and field experiments
- The weight attached to data components
- The assumptions in the input data for a model calculation. The statement on input preparation is needed in cases where the database variables contain insufficient information to perform the calculation. The variables taken from the database in order to generate input data are called the independent variables. If the independent variables available from the database are not in the form required by the model, then some modification of the model or some preprocessing of the input may be required.

**The Model Characteristics.** A selection should be made of the variables to be compared with the output data of the model. These variables are called dependent variables. Some consideration and explicit documentation is required concerning the selection. In heavy gas dispersion modeling, the application is commonly the description of hazard effects (toxicity, flammability) for use in risk analysis. The validation attaches weights to specific variables in a manner that depends on the model characteristics. Variables that are directly relevant to the intended (possibly restricted) use of the model should be weighted more heavily than those of peripheral interest.

**An Estimation of the Model Uncertainty.** With regard to the model, a specific quantitative assessment is required as to the uncertainty in the input data and the output data. The following are some sources of uncertainty in the output results.
• In numerical models, the grid size and time step have a direct effect on the convergence and accuracy of the output data.
• The modeling of stochastic processes (e.g., turbulence) may also cause deviations in the output data.
• The model’s scientific basis may contain erroneous assumptions.
• The effect of the uncertainty in the input data on the output data should be estimated.

An Estimation of the Uncertainty in the Data. With regard to the data sets in the database, a specific quantitative assessment is required as to the uncertainty in the dependent and independent variables. Whenever possible, this information should be contained in the database description.

Use of Code Comparison Exercises. Instead of validating a single model, a series of models can be compared with the same data sets. This provides the evaluator with information about the performance of various models. If the models are carefully chosen this can also establish a state-of-the-art reference to test other models.

The Results of the Validation. The validation leads to a series of calculated values (e.g., concentrations) and a corresponding series of observed values. These values can be compared using different statistical techniques. The statistical analysis can lead to a single set of validation measures, such as a geometric mean bias and mean variance. Alternatively, and more informatively, it can be pursued to look for trends in the comparison as a function of relevant parameters such as distance, wind speed, and density of the released material.

As well as these quantitative measures, it is sometimes informative to produce scatter plots of predicted versus observed values or to plot the agreement between model and observations as a function of relevant parameters.

Conclusions That May Be Drawn. Evaluators can draw their own conclusions from the statistical analysis results of the validation procedure. Alternatively, they can just present the results of the validation procedure and leave it to the interested party to select the best model for a specific application.

In view of the relative scarcity of experimental data in relation to the large number of parameters that play a role in dense gas dispersion, statistical measures do not have an absolute meaning in the sense that by changing the validation data set, the statistical measures can change significantly. Therefore, statistical performance measures of different models can only be compared if exactly the same data sets are used for all models.

Recommendations. The evaluator can close the validation with recommendations on model improvements or experiments for extending the database.

19.6 EXAMPLES OF EVALUATION EXERCISES

19.6.1 Study by Hanna et al.

The extensive work of Hanna et al. (1991, 1993) is the first model evaluation project in which standard objective means of validating heavy-gas dispersion models were applied. The performance evaluations were based on a modeler’s data archive, which contains data sets from eight field experiments. This allowed Hanna et al. to obtain conclusive results on the performance of 14 dispersion models. Six models were ranked as “better” models. These models have their concentration predictions within a factor two of observations in 70–80%
of all cases. There appeared to be no correlation between the quality and complexity of the models.

### 19.6.2 REDIPHEM

The REDIPHEM project was aimed at two aspects of model evaluation. First, a database system was designed and implemented to make experimental data sets available that are useful for evaluation dense gas dispersion models (Nielsen & Ott, 1995). This database has proven to be an excellent platform for dissemination of experimental data related to dispersion. In parallel, the general MEG evaluation protocol (EC, 1994a) was modified to specifically address dense gas dispersion models. A limited, by no means comprehensive, attempt was made to perform a scientific assessment of a number of models developed or improved under support from the European Commission’s research programs (Bakkum et al., 1996).

### 19.6.3 Heavy Gas Dispersion Expert Group

The Heavy Gas Dispersion Expert Group (HGDEG), which was set up by the European Commission’s Model Evaluation Group, developed an evaluation methodology based on the REDIPHEM activities and tested it through a small evaluation exercise. Cole and Wicks (1995) give details and experience of this limited exercise. The objective was to develop a protocol for evaluation of the models and make an overview of the experimental data available for such an evaluation. It was concluded that a more extensive evaluation exercise is needed and that the scientific evaluation requires more attention (Mercer et al., 1998; Duijm et al., 1997).

### 19.6.4 SMEDIS

Continuing the activities from the HGDEG, SMEDIS is an ongoing research project funded under the European Commission’s Environment and Climate Research and Technical Development Program. Its main objective is to develop a methodology for the scientific evaluation of dense gas dispersion models and to test this methodology by actually carrying out the scientific evaluation of a large number of models currently available in Europe. The project is focusing on situations in which complex effects such as aerosols, topography, and obstacles are important as well as “simple” situations. The Health and Safety Executive (HSE, UK) coordinate the project. Thirteen partners cooperate in the project.

SMEDIS is the first project to combine scientific assessment with validation against observed data, including these complex effects, and to apply its procedure to a large number (almost 30) of dense gas dispersion models—the majority in use across Europe. Furthermore, the goal of SMEDIS is to encourage continual model improvement, rather than to rank a set of models at one instant in time, by leaving in place a protocol and archived database of test cases, which can be used by all DGD developers and users in the future.

Daish et al. (1999) have given a general description of the entire project. The scientific assessment is based on a further specification of the protocol and the issues described in this chapter. Carissimo et al. (2001) present results of the first phase of the validation part of the exercise. In this publication models are grouped together according to their type (work book, integral models, shallow-layer models, or CFD), so the performance of individual models is hidden. Results are presented for different release scenarios, covering dispersion over flat terrain without complex effects, dispersion close to obstacles, dispersion of aerosol clouds, and dispersion over complex terrain (see Table 19.8). Contrary to the study by Hanna et al. (see Section 19.6.1), improved performance is observed with increasing model complexity.
TABLE 19.8  First Phase of Results from the SMEDIS Study. FA2 Statistics for Arc-wise Maximum Concentrations for Models Grouped According to Type

<table>
<thead>
<tr>
<th>Model Type</th>
<th>No complex effects</th>
<th>Dispersion close to obstacles</th>
<th>Aerosol effects</th>
<th>Dispersion over complex terrain</th>
</tr>
</thead>
<tbody>
<tr>
<td>Work book (Phenomenological Models and Screening Tools)</td>
<td>40%</td>
<td>42%</td>
<td>43%</td>
<td>Not applicable</td>
</tr>
<tr>
<td>Integral models (box models, steady plume models, and integral plume models)</td>
<td>74%</td>
<td>79%</td>
<td>69%</td>
<td>33%</td>
</tr>
<tr>
<td>Shallow-layer models</td>
<td>65%</td>
<td>53%</td>
<td>32%</td>
<td>50%</td>
</tr>
<tr>
<td>Computational fluid dynamics models</td>
<td>No validation runs performed</td>
<td>89%</td>
<td>75%</td>
<td>71%</td>
</tr>
</tbody>
</table>

Source: Carissimo et al., 2001.

on average, although this table does not exclude that some “simple” models perform as well for some scenarios as more complex models.

19.6.5 Accident Phenomenological and Consequences Assessment

The U.S. Department of Energy (DOE) established the Accident Phenomenology and Consequence (APAC) Methodology Evaluation Program to identify and evaluate methodologies and computer codes to support accident phenomenological and consequence calculations for both radiological and nonradiological materials at DOE facilities and to identify development needs. The program did not include validation, but some models were selected to run test problems. One hundred and thirty-five models were considered and 24 models were finally included in the evaluation (this program did not specifically aim at dense gas dispersion). The scientific assessment was based on the following 10 model attributes or features:

1. Source term algorithm
2. Input parameters required to run the transport and dispersion model
3. Dispersion submodel type
4. Model capabilities/physics
5. Transport submodel
6. Meteorological input
7. Health consequences submodel
8. Source/receptor mitigation measures
9. Output capabilities
10. Uncertainty analysis

It is recommended that the study be completed with a validation exercise (Lazaro et al., 1997).


