CHEMICAL ENGINEERING TRANSACTIONS
VOL. 116, 2025
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Bruno Fabiano, Valerio Cozzani

ISBN 979-12-81206-18-2; **ISSN** 2283-9216
DOI: 10.3303/CET25116009

Using a Surrogate Model in Risk Studies Using CFD Simulations

Shona Mackie, Connor Bloodworth, Chris Coffey

Gexcon AS, Fantoftvegen 38, NO-5072 Bergen, Norway
shona.mackie@gexcon.com

Risk is a combination of the frequency with which consequences occur, and their cost. We calculate it by estimating the range and frequencies for different initial conditions and calculating predictions that cover these to achieve frequency distributions of consequences, which we combine with loss and fatality models to account for cost. This requires calculation of enough consequence predictions to cover all realistic possibilities, which is often computationally impractical using computational fluid dynamics (CFD) simulations. We investigated using a surrogate model to cheaply calculate reliable consequence predictions for tens of thousands of scenarios. The surrogate is built using the inputs and outputs from CFD simulations. We investigated different methods of selecting scenarios for these simulations, and different methods of building a surrogate model. An approach based on a Gaussian process model and a recurrent neural network resulted in predictions of maximum equivalent cloud volume and a predicted probability density function for Q9 volume that are similar to the CFD predictions. The surrogate-predicted distribution is wide enough to capture less frequent events (that are often higher-cost) and can be combined with loss and fatality models to calculate a risk assessment. We used the FLACS-CFD model to demonstrate the method for an example case, focused on a dispersion scenario with a single leak and predictions for equivalent cloud volume. The method could be implemented for other scenario types and other predictions, and for other CFD models. The results are promising, suggesting that this type of surrogate may be useful for risk studies, where it is helpful to have many more predictions than it is computationally practical to simulate with CFD.

1. Introduction

When assessing risks for an installation, it is insufficient to consider consequences from only one scenario. Initial conditions may vary and the consequences of an incident under any combination of these should be considered in a risk study. It is generally impractical to run thousands of simulations using CFD (to predict consequences from thousands of slightly different scenarios) and this raises issues:

- Limiting the number of initial conditions represented in the study means that not all risks are considered.
- Different experts make different, equally justified, choices for which scenarios to simulate, leading to different risk assessments for the same case.
- It is possible, but not always appropriate, to interpolate between simulated consequences (and it is not easy to discern whether or not it is appropriate for a given case).
- The sensitivities of CFD calculations to different inputs are not independent, so one-dimensional interpolation is generally inappropriate.
- CFD models are non-linear, which may lead to local consequence peaks in some areas of the input space. Such peaks are likely to be missed if one-dimensional interpolation is used.

A surrogate model, built for the geometry and conditions of a specific case, can estimate predictions at trivial computational expense. Where a project has a budget of 1000 CFD simulations, it may be beneficial to use those to build a surrogate, from which tens of thousands of predictions can be calculated. This may result in an accurate range of predictions that covers the range of initial conditions more thoroughly than 1000 direct simulations could.

Paper Received: 22 November 2024; Revised: 29 January 2025; Accepted: 22 April 2025
Please cite this article as: Mackie S., Bloodworth C., Coffey C., 2025, Using a Surrogate Model in Risk Studies Using CFD Simulations, Chemical Engineering Transactions, 116, 49-54 DOI:10.3303/CET25116009

Our goal is to implement this for a realistic case and to assess the reliability of the resulting predictions. Ideally, we hope for more information at lower computational expense with no reduction in accuracy.
Our prototype selects scenarios to simulate, runs CFD simulations, uses the CFD-output to build a surrogate model, and then uses the surrogate model to calculate predictions. The result is a distribution that represents the estimated frequency of the predicted consequence. It is straightforward to test whether enough predictions have been calculated by increasing the number of predictions and comparing the resulting prediction distribution. If enough predictions have been calculated, then their distribution will not change as the number of predictions increases. The frequency distributions calculated using the surrogate model can be combined with loss and fatality models to account for cost to complete a risk assessment. If inspection of results leads to a requirement for more predictions, then it is trivial to calculate these. We consider simple interpolation methods, since these are similar to the approach commonly taken manually in safety studies, and we also consider a more complex method, combining a Gaussian process model with a recurrent neural network. This latter approach has been used for CFD in wind engineering (Weerasuriya et al, 2021), but generally with fewer varying inputs than we look at here, and without coupling the Gaussian process model to a recurrent neural network. Using a Gaussian process model with a recurrent neural network has been applied successfully for timeseries predictions in other fields, such as battery health and Mackey-Glass timeseries forecasting (Sun et al., 2022), and to model timeseries data with multiple parameters for detecting serious health conditions (Futoma et al, 2017). This suggests it may be an effective surrogate approach for a typical CFD-based risk study.

2. Example case

We anticipate a surrogate be useful when dispersion simulations are used to predict a distribution of cloud properties for later explosion simulations. Our example case is a release from a point-leak of light gas (99% methane, 1% ethane) in a congested region in offshore platform in Figure 1. The leak position and orientation were varied along with other initial conditions, and we looked at equivalent cloud volume (Q9 volume) predictions.

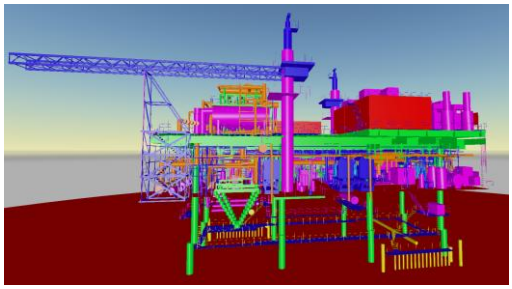

Figure 1.Small offshore platform used as an example case.

2.1 Initial conditions

Initial conditions were varied following realistic assumptions for a risk study:

- Wind speed: Gaussian distribution, centred on 9 ms^{-1}, with standard deviation 3 ms^{-1}.
- Wind direction: Gaussian distribution, centred on 270° from North, with a standard deviation of 60°.
- Pasquil classes: D, E and F.
- Ambient temperature: Gaussian distribution, centred on 20°C, with standard deviation 5°C.
- Ambient pressure: Gaussian distribution, centred on 10 0000 Pa, with standard deviation 1500 Pa.
- Leak location: varying uniformly in X between 98.5 and 115.5 m; in Y between 96.0 and 102.0 m, and in Z between 11 and 16 m (coordinates give position relative to the origin in the scenario geometry).
- Leak direction: All directions.
- Reservoir pressure: Gaussian distribution, centred on 145 bar, with standard deviation 0.725 bar.
- Reservoir temperature: Gaussian distribution, centred on 70°C, with standard deviation 0.35°C.
- Leak diameter: Uniform distribution between 0.01 m and 0.03 m.

3. Code structure

Our prototype is written in Python3 and uses FLACS-CFD software, and the method is outlined in **Errore. L'origine riferimento non è stata trovata.**.

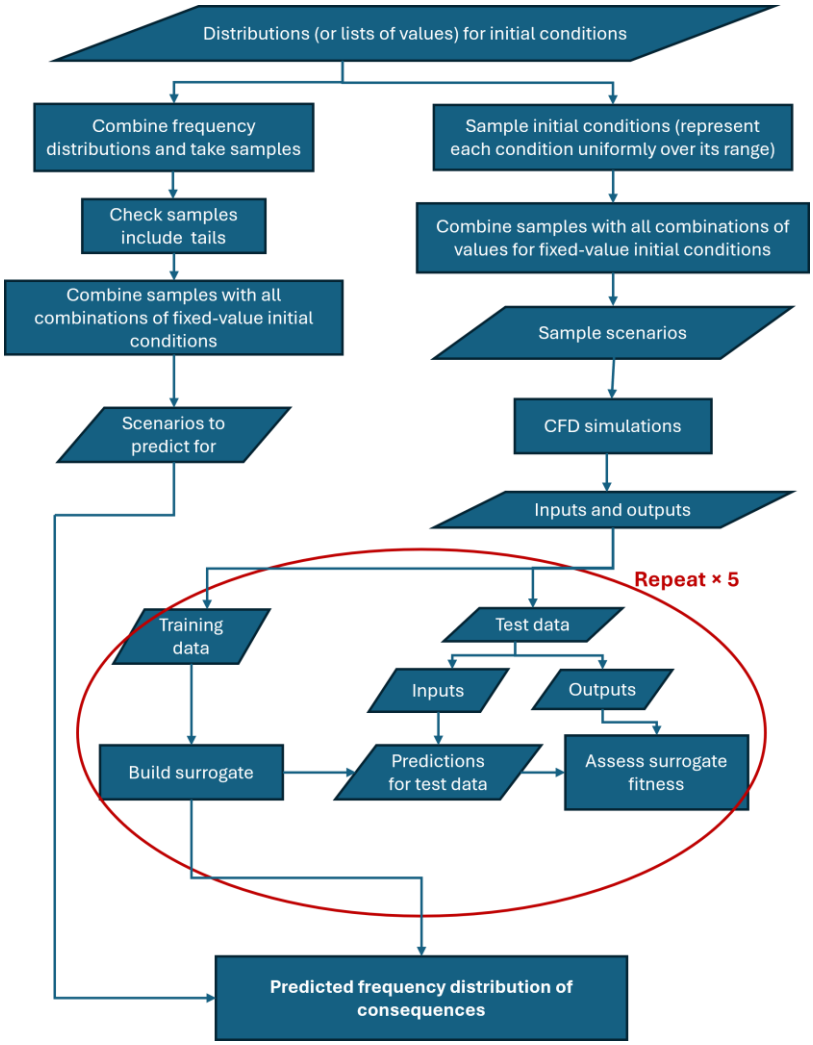

Figure 2. General steps to build, test and use surrogate model.

3.1 Input script

A user-provided input script specifies which initial conditions should be varied, and how, and sets the maximum number of CFD simulations to run. Initial conditions that vary are specified as a list of fixed values or as a distribution. If fixed values are supplied, then predictions are calculated for each value and no other values will be considered for that parameter. For example, if 3.2, 4.5 and 6.7 ms^{-1} are specified for wind speed, then predictions for these, and no other, wind speeds will be calculated. This may be useful when a finite set of possibly values is known for an initial condition. Alternatively, a distribution can be provided, and must be either Gaussian, lognormal or uniform. The latter is defined by its minimum and maximum values and the former two are defined using mu and sigma, and (optionally) minimum and maximum limits for truncation. Distributions for wind speed, relative turbulence intensity, leak area, mass flow, leak velocity, reservoir volume, characteristic velocity and leak diameter are automatically truncated with a lower bound of 0 if an unbounded distribution is specified. If the Pasquil class is varied, then it must be supplied as a list of values.

4. Selecting data to build a surrogate

The range of values given in the input script for each initial condition is used to create a uniform distribution. These are combined and the combined distribution is sampled from to create scenarios for simulation. The simulations then provide the data to build the surrogate. If an input is provided as a list of fixed values, then every set of inputs sampled from the combined distribution is simulated with each fixed value. If multiple inputs are specified as a list of fixed values, then all possible combinations of these are considered. We assumed a

budget of 1000 simulations and so created 999 sample simulations (the number of simulations must be divisible by the number of combinations of values for the fixed-value inputs). We selected the samples using Sobol sampling, a multi-dimensional technique that covers the parameter space thoroughly and is extensively used for building emulators to represent complex models (Saltelli et al., 2002).

5. Building a surrogate model

The relative sensitivities of a CFD model to individual varying initial conditions (model inputs) are likely to depend on the specific scenario being modelled, i.e., to depend on the non-varying initial conditions, and at least some of the sensitivities are likely to be related. For a method to be effective for CFD safety studies, it must handle multiple inputs (so that variations in multiple initial conditions are considered together) and must be generic enough to be applicable to different complex non-varying initial conditions, such as geometry (a common CFD application is modeling gas dispersion from a leak on a complex structure such as an offshore platform, leading to a fire or an explosion).

5.1 Interpolation methods

We built a nearest-neighbour interpolator and two radial basis function (RBF) interpolators, using a linear and a cubic kernel, to interpolate between the CFD-simulated predictions for the given inputs. RBF interpolation looks at the distance between the interpolated point and every simulated point in every dimension, so it interpolates with respect to all initial conditions simultaneously and considers all simulated predictions when estimating each single prediction.

5.2 Gaussian process-recurrent neural network model (GPR-RNN)

In contrast to the predictions from the interpolation methods, which provide a single Q9 volume prediction per timestep, the GPR-RNN predicts a probability distribution of Q9 volumes for each time step. The uncertainty associated with the surrogate method is therefore implicitly captured in the surrogate predictions. Under this hybrid approach, the timeseries data from the CFD output were split into normalized 1D sequence data and a scale factor (the scale factor is the maximum Q9). The scale factor is estimated using Gaussian process regression (GPR), a non-parametric Bayesian approach that provides a probabilistic prediction. GPR models the underlying function as a Gaussian process and uses a covariance kernel to determine the similarity between the CFD-simulated predictions, allowing it to provide a mean prediction and variance for each point in the input space where a surrogate prediction is required. This method is particularly powerful for modelling complex, non-linear relationships and providing uncertainty estimates. The sequence data are estimated using recurrent neural networks (NN). These have a unique architecture that allows information to persist across steps in a sequence, making them appropriate for tasks involving sequential data. The uncertainty around each point in each sequence was estimated by analysing the residuals from a comparison between the predictions and the CFD-simulated values, for a set of inputs that were not included in training the NN. The distribution of predictions, combining the scale and sequence data, was then calculated by using numerical methods to determine the product of each sequence point with the distribution for the predicted maximum Q9 volume. For details of the GPR method followed here, see Rasmussen and Williams, 2006.

6. Surrogate fitness

It is essential that a surrogate model reliably reproduces the results of the full model. We created five test datasets by selecting 150 of the simulations using random sampling with a fixed seed, which we varied five times. We built the surrogates using the remaining simulations and compared the surrogate- and CFD-predictions for each test dataset.

6.1 Surrogate fitness for maximum Q9 predictions

We compared the surrogate- and CFD- predictions for maximum Q9 in **Errore. L'origine riferimento non è stata trovata.**. The GPR-RNN estimates match those the CFD predictions most closely and the 1:1 line passes through the uncertainty bounds for many points that otherwise appear scattered in **Errore. L'origine riferimento non è stata trovata.**. This shows that using the distribution for each prediction, rather than just the mean for each prediction, gives a more accurate estimate of maximum Q9 volume.

6.2 Surrogate fitness for Q9 volume predictions

We plotted the probabilities of exceedance for Q9 volumes, as predicted by the surrogate and by CFD in Figure 4. Results are shown for GPR-RNN only, since this provided the closest match to the CFD predictions. We also compared the similarity of the probability density functions (PDFs) for Q9 volume that were predicted by the

surrogate models and by the CFD simulations. Space constraints mean that these are not shown here, but the Jensen-Shannon divergence (JSD) is displayed on the plots in Figure 4.

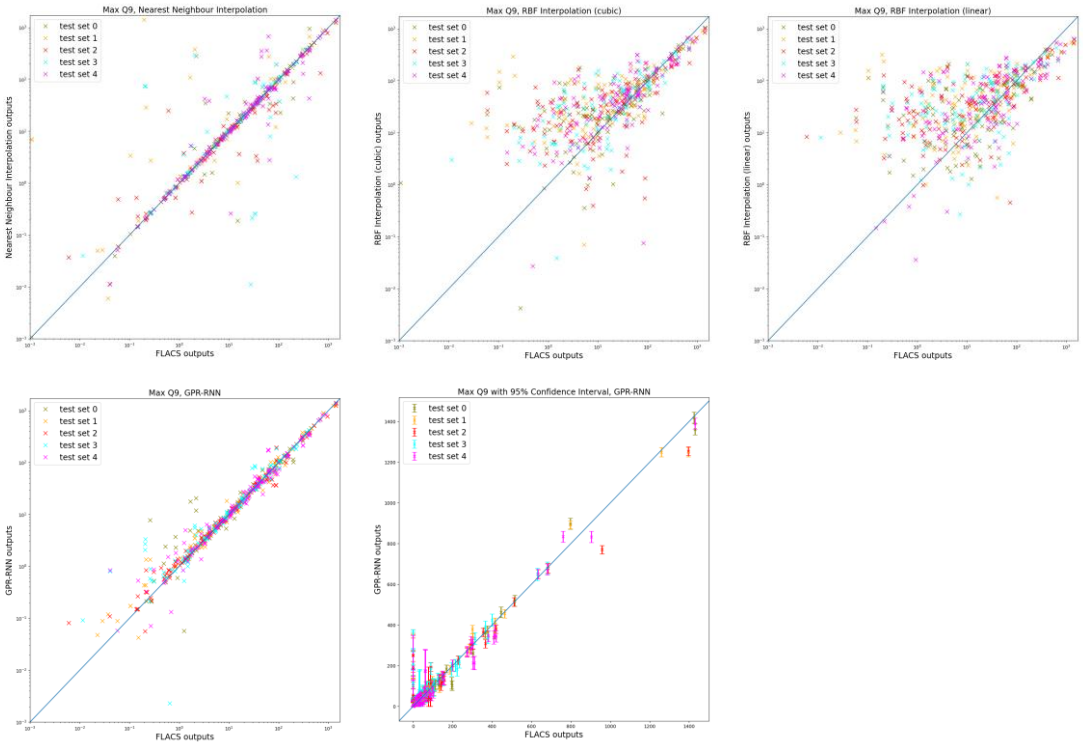

Figure 4. Maximum Q9 volume predictions from CFD (x-axes) and from each surrogate (y-axes). GPR-RNN gives a distribution for each maximum prediction at each timestep, the maximum in the lower left plot was calculated from the means of these. The lower right plot shows the same information, but on linear axes with ±1.96 sigma error bars.

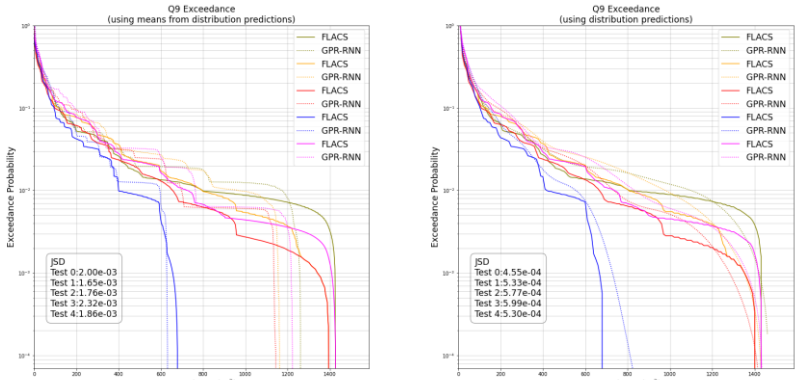

Figure 3. Probabilities of Q9 volume exceedance predicted by CFD (solid) and by GPR-RNN (dotted). The JSD is shown as in insert in the plots as a measure of similarity between the predicted Q9 volume PDFs (not shown), see text for an explanation of JSD. Left: results from the mean of each time-step prediction from GPR-RNN; Right: results from aggregating the distribution of GPR-RNN predictions for each timestep.

We calculated the JSD between the log-transformed PDFs to ensure that the metric captured differences between the PDF tails, which represent high Q9 volumes (JSD would otherwise reflect differences around the PDF peaks more strongly). GPR-RNN provides predictions as Gaussian distributions, and we calculated the Q9 volume exceedances and PDFs using the means for each GPR-RNN prediction, and by aggregating the

predicted Q9 volume distributions for each timestep. This latter approach results in broader PDFs that more closely match those from CFD (giving a lower JSD), and exceedance probabilities that are closer to the CFD predictions, as seen in Figure 4.

7. Final Predictions

We take 10 000 samples from the combined distribution of the provided initial-condition distributions and implement a check to ensure that the tails are included (low-frequency events are often associated with high costs). For initial conditions that were provided as a list of values, each possible combination is combined with every sample from the joint distribution. The surrogate predictions are shown in Figure 5.

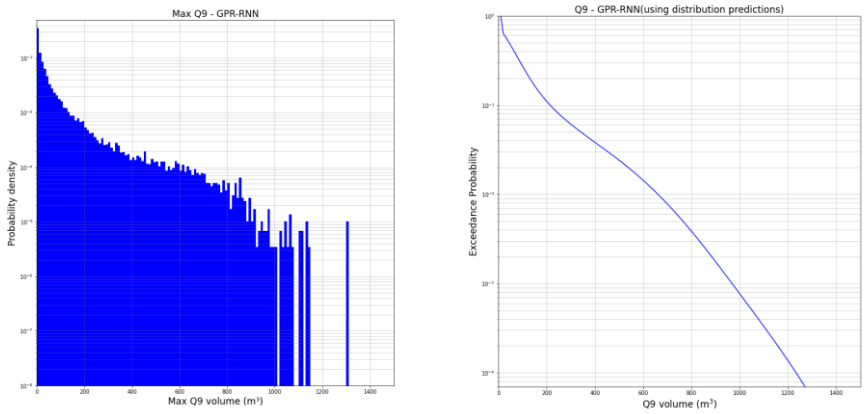

Figure 5. Surrogate predictions: distribution of max Q9 volume (left); probability of Q9 volume exceedance (right).

8. Conclusions

This demonstration shows that a GPR-RNN surrogate has the potential to provide computationally cheap and accurate predictions for CFD calculations, making it practical to consider more scenarios in a risk study, and so reducing the need for subjective scenario selection. It considers variability in all parameters simultaneously and its intrinsic fitness measure shows how reliably its predictions reproduce CFD results (the surrogate results share any biases present in the CFD predictions). The large number of simulations needed to train the surrogate offsets the computational saving and it may provide little saving for studies where only one or two initial conditions are varied. However, when multiple initial conditions are varied, then the huge (in principle, unlimited) increase in the number of scenarios that can be considered make this a powerful method of reducing computational cost without compromising accuracy.

References

Futoma, J. Hariharan, S., Heller, K., Sendak, M., Brajer, N., Clement, M., Bedoya, A. and O'Brien, C. 2017. An Improved Multi-Output Gaussian Process RNN with Real-Time Validation for Early Sepsis Detection. Proceedings of Machine Learning for Healthcare 2017.

Rasmussen, C. E. and Williams, C. K. I. 2006. Gaussian Processes for Machine Learning. The MIT Press.

Saltelli, A., Tarantola, S., Campolongo, F. and Ratto, M. 2002. A Guide to Assessing Scientific Models. Wiley.

Sun, X., Kim, S. and Choi, J. I. 2022. Recurrent neural network-induced Gaussian process. Neurocomputing 509, p. 75-84.

Weerasuriya, A. U., Zhang, X., Lu, B., Tse, K. T. And Lui, C. H. 2021. A Gaussian Process-Based emulator for modelling pedestrian-level wind field. Building and environment 188.