Uncertainty Quantification for Chemical and Biological Hazard Assessment

Dr Ronni Bowman
Hazard Assessment, Simulation and Prediction Group

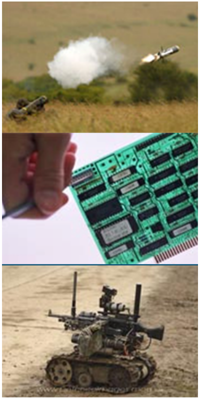
vbowman@dstl.gov.uk

Outline

- Background
- Hazard Assessment and Dispersion Modelling
- Methodology
  - Sensor Placement
  - Source Term Estimation (STE)
  - Hazard Chain
- STE in Detail
- Emulation
- Uncertainty Calculation
- Uncertainty Presentation

Defence Science and Technology Laboratory (Dstl)

- MOD's science and technology experts.
- Provide independent, impartial S&T advice to MOD and UK government.
- Not just home based. Scientists deployed to support operations.
- Work with very small companies to world-class universities, huge defence companies, government departments and other nations.
- Deep and widespread research for immediate and future requirements.
- Trading fund.



Dstl's Purpose

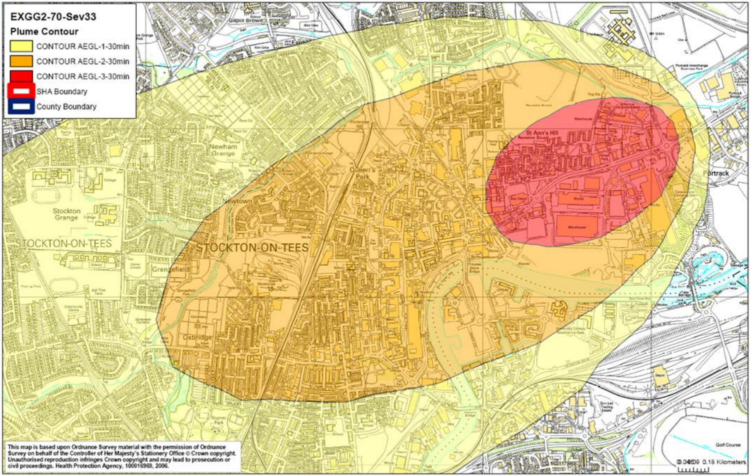
To maximise the impact of science and technology for defence and security of the UK.

- Supply sensitive and specialist science and technology services for MOD and wider government.
- Provide and facilitate expert advice, analysis and assurance to aid decision making.
- Lead the formulation, design and delivery of a coherent and integrated MOD science and technology programme.
- Manage and exploit knowledge across the wider defence and security community.
- Act as a trusted interface.
- Champion and develop science and technology skills across MOD.

Hazard Assessment

- In an emergency involving an accidental or deliberate release of a Chemical or Biological (CB) substance there is an urgent need for a hazard assessment.
- This assessment is delivered in the form of a hazard area, which details areas of contamination at known levels of risk
  - Lethality / Incapacitation / Miosis,
  - Probability of infection (Biological).

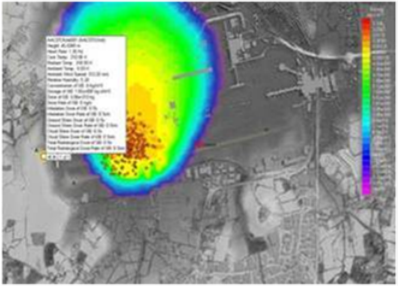
An Example Hazard Area



Dispersion Modelling

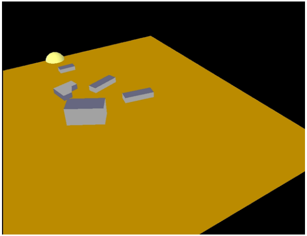
Aim: Prediction of the downwind hazard generated by a chemical or biological (or other) release.

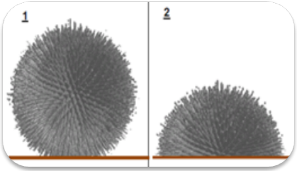
- Accident response; military planning; volcanic ash; ...
- Variety of models
  - Gaussian plume (Clarke, 1979)
  - Gaussian puff (Sykes et al., 1998)
  - NAME (Jones et al., 2007).
- Underpinning capability for the HASP group.



Dispersion

- A CB hazard disperses in the atmosphere and the hazard area is determined by
  - Source Term (dissemination device, mass, efficiency)
  - Meteorology
  - Terrain
  - Building Interactions.



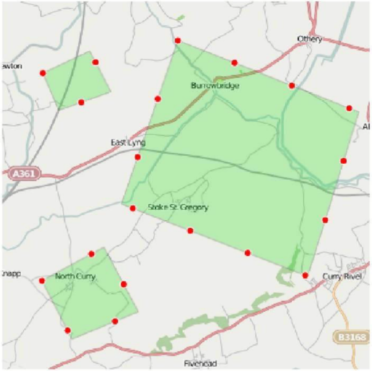


Uncertainty

- Dispersion is highly uncertain and outputs need to be translated into effective information, this requires source inversion and optimization.
- Uncertainty must be propagated in order to provide a complete answer.
- Uncertainty must be represented in a way that is understandable to a military commander.
- It has to be useful, if the uncertainty is too large it could be ignored irrespective of the validity of the calculations.

Sensor Placement

- Tool developed to aid in the deployment of CB assets.
- Tool uses a sample of potential releases and creates a database for optimization
  - Probability of detection
  - Warning time
  - Distribution of assets across areas of the battlespace
  - Desirability of placement.
- Current research into data storage, optimization and dependency.
  - Ideally the tool would rapidly optimize for casualties, however, this an open problem.



Source Term Estimation

- Source term estimation is a highly uncertain inverse problem.
- A source term estimation model has been developed in order to infer CB source parameters from sensor readings (Robins et al. 2009).
- Inference is made by hypothesizing potential releases and calculating their likelihood based on sensor readings and meteorology.
- This likelihood is then combined with various prior distributions to produce a posterior estimate of the likely source term distribution.
- This posterior distribution is then sampled to produce a hazard estimate of where contamination is likely, based on the available data.

Source Term Estimation

- Existence of a release is given a prior via a surrogate mass parameter, m^*, all sampled parameter sets with $m^* \leq 0$, denote no release with $m = 0$. The prior on the surrogate mass is as follows:

$$p(m^*) = \frac{1}{2\mu_{m^*}} e^{-\frac{|m^*|}{\mu_{m^*}}}.$$

  The mean μ_{m^*} is determined according to operational information. If $m^* \leq 0$, the other parameters are maintained but irrelevant to the inference.

- Meteorology must also be inferred due to the uncertain nature of the local meteorological data.

Source Term Estimation

- Likelihood is calculated using sensor readings via the dispersion model.

$$F(c|\mu,\sigma) = \begin{cases} 0 & c < 0 \\ \Phi\left(\frac{c-\mu}{\sigma}\right) & c \geq 0 \end{cases},$$

  where Φ is the standard normal distribution function, c is concentration and μ, σ are the mean and variance of the concentration produced from the dispersion model.

- Source parameters are multi-dimensional and contain 15 parameters including location, time, mass, u and v components for meteorology, surface components and agent.

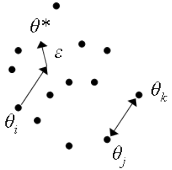
- Meteorological parameters are inferred via readings provided to the system in a similar ways to CB sensor readings.

Source Term Estimation - Proposals

- Proposals via Differential Evolution Markov Chain (DE-MC): Given M chains, new hypotheses update each chain end θ_t^i, for $i = 1, ..., M$.
  1. Select θ_t^i;
  2. Randomly select 2 additional chain ends, (θ_t^j, θ_t^k) where $j, k \neq i$;
  3. Sample $\epsilon \sim S$;
  4. Propose:

$$\theta^* = \theta_t^i + \gamma \left(\theta_t^j - \theta_t^k \right) + \epsilon$$

  $S = N\left(0, \sigma^2\right)$ for small σ, and γ is a multiplication factor that restricts 'step size'.



Source Term Estimation - Computation

- Posterior sampling is complex due to the large number of parameters and the ‘witches hat’ form of the posterior across these dimensions.

- Posterior computation undertaken using a bespoke algorithm based upon Sequential Monte Carlo (SMC) and Sample Importance Resample (SIR):
  - Update weights of each hypothesis
  - Normalise weights so total weight is equal to number of samples N_{eff}:

$$N_{eff} = \frac{\left(\sum_{i=1}^{N} w_i\right)^2}{\sum_{i=1}^{N} w_i^2}$$

  - Resample according to weights.

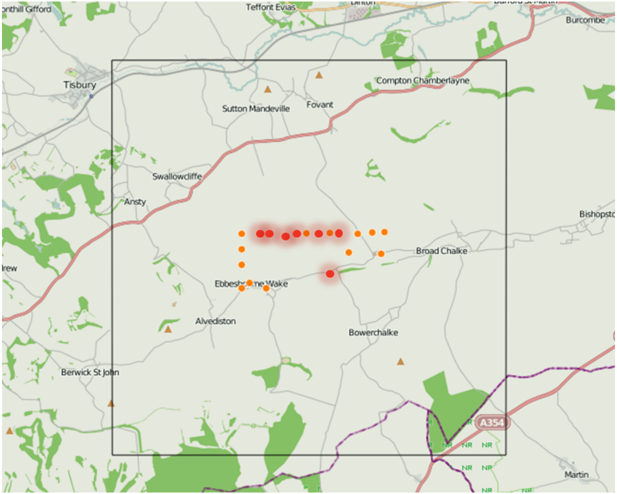
Source Term Estimation - Sampling

- An operational system must determine if a hazard is present, the probability of release is calculated as

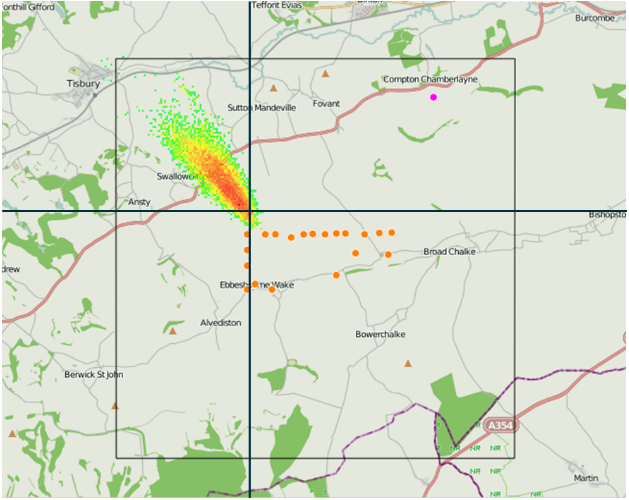
$$P\left(m^* > 0|\mathrm{D}\right) = \frac{P\left(m^* > 0\right)\sum_{ij} w_{ij} I\left(m_i^* > 0\right)}{P\left(m^* > 0\right)\sum_{ij} w_{ij} I\left(m_i^* > 0\right) + \left(1 - P\left(m^* > 0\right)\right)\sum_{ij} w_{ij} I\left(m_i^* \leq 0\right)}$$

  - $P\left(m^* > 0\right)$ is the prior probability of release
  - w_{ij} is the j^{th} weight of the i^{th} hypothesis
  - $I\left(m_i^* > 0\right)$ is an indicator function that returns one if the hypothesized mass is strictly positive (i.e. $m^* > 0$; a possible release) and zero otherwise (no release).

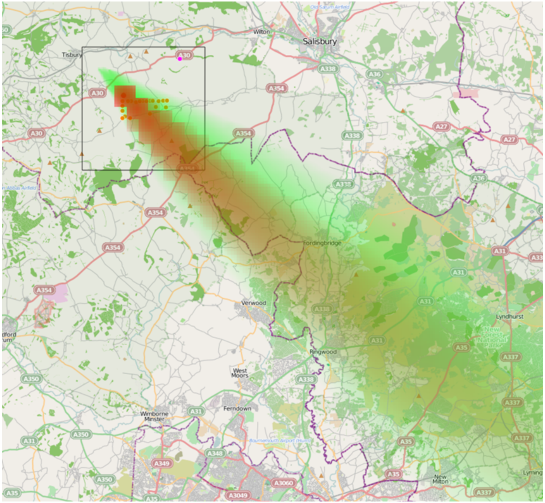
Source Term Estimation - Sensor Alarm



Source Term Estimation - Inference



Source Term Estimation - Prediction

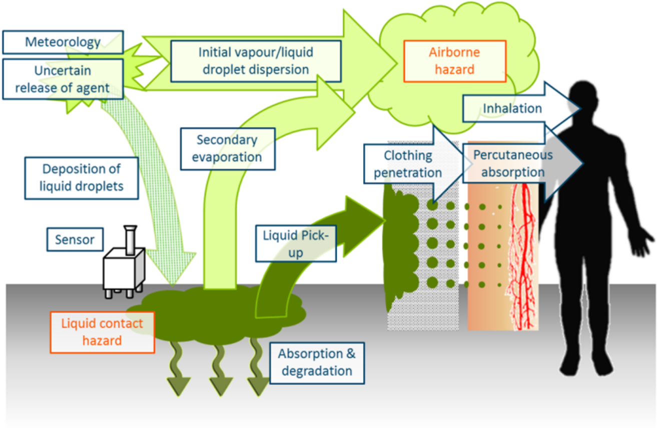


A Hazard Chain

- A hazard area is calculated in a number of separate steps:
  - Inputs
    - Meteorology
    - Source - Mass, Release mechanism etc.
  - Dispersion Model
  - Dose Calculation
  - Dose response curve.

- Each step is complex with numerous inputs and model choices.

A Hazard Chain



Uncertainty Quantification

- Currently the biggest limitation to accurate hazard prediction.
- Each step in the modelling chain has inherent uncertainty from numerous sources.
- Naively, simulation studies could be used to understand the uncertainty in predictions
  - Typically models are too computationally expensive
  - Model inaccuracies must also be accounted for
  - Statistical models must be combined with real data at differing points in the modelling chain.
- Each step is time consuming, however, answers are required in real time.
- Uncertainty must be communicated effectively.

Emulation

- An initial study into the potential use of emulators focused on the emulation of the underpinning dispersion model.
- Research suggests that while emulation is possible there are significant challenges:
  - Input parameters can result in significantly different functional output
  - Output is functional but also in several different forms
  - Meteorological and terrain constraints may require an emulator to be developed for each location.

Multivariate Emulation

- Let $\boldsymbol{x}_i = (x_{1i}, \ldots, x_{q_1 i})$ be the vector of input values at which the ith run of the simulator is performed.

- Let $\boldsymbol{Y}_i = (Y_1(\boldsymbol{s}_1), \ldots, Y_r(\boldsymbol{s}_r))^{\mathrm{T}}$ be the vectorised output from this run .

- The vector $\boldsymbol{s}_j = (s_{1j}, \ldots, s_{q_2 j})$ locates the jth output in the q_2 dimensional output domain.

- Dimension reduction is obtained through assuming, for each output vector, the linear model

$$\boldsymbol{Y}_i = \sum_{k=1}^{p} \boldsymbol{a}_k(\boldsymbol{s})\beta_k(\boldsymbol{x}_i) + \boldsymbol{e}_i\,.$$

Multivariate Emulation

- Linear Model

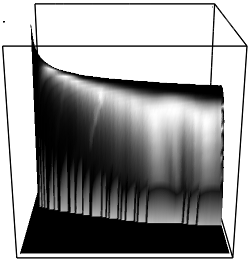
$$\boldsymbol{Y}_i = \sum_{k=1}^{p} \boldsymbol{a}_k(\boldsymbol{s})\beta_k(\boldsymbol{x}_i) + \boldsymbol{e}_i \,.$$

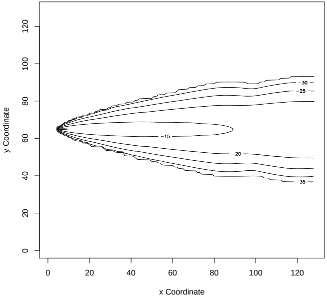
- Here $\boldsymbol{a}_1(\boldsymbol{s}), \ldots, \boldsymbol{a}_p(\boldsymbol{s})$ are a set of $r \times 1$ basis vectors which are assumed independent of $\boldsymbol{x}_i$ but which may depend on the indexes $\boldsymbol{s} = (\boldsymbol{s}_1^{\mathrm{T}}, \ldots, \boldsymbol{s}_r^{\mathrm{T}})^{\mathrm{T}}$.

- The corresponding coefficients $\beta_1(\boldsymbol{x}_i), \ldots, \beta_k(\boldsymbol{x}_i)$ may depend on the inputs $\boldsymbol{x}_i$, and $\boldsymbol{e}_i$ is a r-vector of errors resulting from the basis function approximation.

- Let $\boldsymbol{\beta}(\boldsymbol{x}_i) = (\beta_1(\boldsymbol{x}_i), \ldots, \beta_p(\boldsymbol{x}_i))^{\mathrm{T}}$.

Application to Dispersion



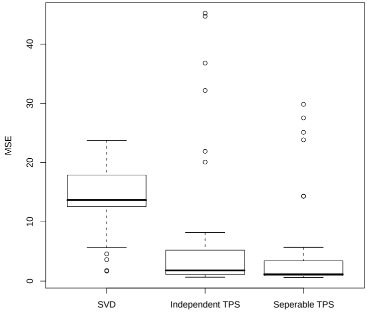


A Typical Dosage Output from the Dispersion Model Output on a Log Scale.

Emulation Approaches

- Three emulation approaches are applied and compared
  1. A fully Bayesian approach using a principal components basis (PC emulator; Higden et al. 2008).
  2. A fully Bayesian approach using a thin plate spline basis (Wood (2003)) and assuming independence of the elements of $\boldsymbol{\beta}(\boldsymbol{x}_i)$ (Independent TPS emulator).
  3. A "plug-in" Bayesian approach using a thin plate spline basis and assuming a separable covariance structure (Rougier(2008)) for $\boldsymbol{\beta}$ (Separable TPS emulator).
- Posterior predictive distributions for emulators 1 and 2 are obtained via MCMC and $W^s(\boldsymbol{s}) = I_p$, within run correlations are assumed to be independent - overconfidence can result in emulator 2.
- The posterior for emulator 3 is obtained via a plug-in approach.

Emulator Comparison



Mean squared errors for each of the PC, Independent TPS and Separable TPS emulators calculated using the posterior predictive mean across the test set.

Communication

- The overall Hazard area is highly uncertain, however information must be conveyed in a concise and clear manner for decision makers.
- Large uncertainties can be counterproductive - a course of action must be obvious.
- Underestimation of the hazard area could have severe consequences and must be avoided (over-estimation is far more acceptable within the bounds above).
- Spatial uncertainty is difficult to portray and this is an open problem.

Conclusions

- Hazard assessment is a complex problem involving:
  - Multi-objective optimization of large multi-dimensional data sets.
  - Source Inversion under complex meteorological conditions in real time.
  - Propagation of uncertainty through highly complex modelling chains in real time with multiple uncertainty types.

- There are tools under development, however, the concatenation of these tools and their enhanced development are open problems.
  - A method of optimization over a multi-objective, multi-dimensional space.
  - An integrated modelling chain capable of source estimation and prediction in real time.
  - Uncertainty propagation in real time through the modelling chain.

Selected references

- Higdon, D., Gattiker, J., Williams, B. J. and Rightly, M. (2008). Journal of the American Statistical Association 103 570-583.
- Jolliffe, I. T. (2002), 2nd ed. Springer, New York.
- Morris, M. D. and Mitchell, T. J. (1995) Journal of Statistical Planning and Inference 43 381-402.
- Rasmussen, C. E. and Williams, C. K. I. (2006). MIT Press, Cambridge, MA.
- Robins, P., Rapley, V. E. and Green, N. (2009). Journal of the Royal Statistical Society C 58 641-662.
- Rougier, J. C. (2008). Journal of Computational and Graphical Statistics 17 827-843.
- Wood, S. N. (2003). Journal of the Royal Statistical Society B 65 95-114.