# Surrogate Modelling of Spray Dispersion using Computational Fluid Dynamics & Machine Learning

Dominik Klotz and Samuel Dale

*Department of Chemical Engineering, Imperial College London, U.K.*

**Abstract** *Spray-mist deposition from vehicles is an important phenomenon in several fields including pesticide application and disinfection of large areas. Here, the feasibility of the use of machine learning (ML) to generate surrogate models of spray behaviour from computational fluid dynamics (CFD) data is investigated. CFD is used to model a single cone spray nozzle using a Lagrangian-Eulerian approach, with pesticide spray modelled as a kinematic particle cloud. A self-generated computational pipeline is used to produce 5550 distinct simulations with variation of multiple physical parameters. A 600 pixel image of spray deposition at each simulation time is produced for each simulation and is used to train various classes of ML algorithms. The strongest surrogate model approaches were found to be based on a self-developed traversing local frame on which a Gaussian process regressor was trained, achieving a maximum root relative mean error of 1.3%, and thresholding of the deposition images with binary images being used for classifier training which showed an overall accuracy of 78%. The potential of surrogate modelling of this type of chaotic flow behaviour was thus shown. This technique is potentially suitable for active control of spray dispersion due to its expected advantages of high accuracy and configurability with low end-user computing times.*

## 1 Introduction

Spray-mist deposition from either terrestrial vehicles, aircraft or, increasingly, unmanned aerial vehicles (UAVs) is the dominant route of pesticide application. Globally, 3.5 Mt of pesticides are projected to have been applied during 2020 [1]. Depending on the mode of pesticide application, up to 75% of the applied volume does not reach the intended target area – the annual cost of environmental and societal damage by unnecessary pesticide exposure is estimated to amount to up 10 billion US dollars in the US alone [1, 2].

In order to apply pesticides, solutions of varying pesticide and additive compositions are sprayed out via boom arrangements of several nozzles; the atomisation process results in broadly distributed spray droplet diameters on the order of magnitude of $\mu$m. Lower droplet diameters achieve better, uniform active component coverage of the plant foliage but are more susceptible to downwind displacement [3]. Horizontal downwind drift of spray droplets outside their intended area of deposition is referred to as spray drift, and is the primary pathway for off-target pesticide deposition [4].

Efforts to mitigate spray drift have lead to positive results, particularly the development of drift-reducing nozzles (DRN) [5], control of atomisation and evaporation properties via additive by-mixtures [6], and introduction of agricultural buffer zones [7]. Accurate monitoring of dispersion effects, and thus the establishment of conclusive cause-effect relationships in legislation, is incredibly difficult due to weather-sensitivity and non-periodic spraying schedules [8]. Regulatory agencies have established various safety requirement schemes, created extensive rating frameworks for various nozzle types and placed heavy weighting on the wide-spread roll out of DRNs [9, 10]. The classification schemes usually rely on comparison to a reference nozzle under specified wind tunnel conditions. Comparative experiments under strictly defined conditions provide adequate results for differentiation of individual nozzle designs and their spray drift potentials, but provide limited information for modelling of actual field spray drift behaviour [3].

With the rise of precision agriculture [11] and smart farming [12], the level of technological capabilities in the agricultural sector is primed to enter a new phase. Application of sophisticated sensor and control techniques via UAVs will enable optimally targeted application of pesticides [13–15] – and the development of easily deployable, highly accurate spray drift models will be required to successfully adapt and control spraying schedules in real time, enabling further reduction of the environmental impact of pesticide usage.

Spray dispersion from UAVs has found wider use in 2020, with efforts to modify agricultural UAVs for disinfectant spraying taking place in response to the COVID-19 pandemic [16] and the rapid roll-out of commercial UAV disinfection services having occurred [17]. Reduction of spray drift is important here to ensure effective disinfection of the desired surface and prevent environmental contamination by dispersed disinfectant.

This report sets out to develop a data-driven surrogate model (SM) of the spray drift phenomenon, ca-

pable of achieving highly accurate results at low computational costs and being useful for real-time control of spray-mist deposition. Development of a validated computational fluid dynamics (CFD) simulation (3.1) and integration into a custom computational pipeline (3.2) allowed for generation of high quality synthetic CFD data used to train multiple classes of ML algorithms to produce surrogate models of spray behaviour (3.3). Analysis showed that the most successful approach was based on image thresholding and classification of individual pixels (4.2.4), with a self-developed traversing local frame regressor also showing promise (4.2.4). The potential of ML-based surrogate modelling of spray behaviour was demonstrated, with potential applications in active control of spray dispersion or in nozzle design studies.

# 2 Background

Currently available spray drift models can be compartmentalised into four main approaches, summarised in order of increasing complexity from (1) to (4) in table 1 [18–23]. Atmospheric plume dispersion and multiple regression are computationally inexpensive, but provide only limited accuracy: they remain extensively used in the assessment of long term exposition and emission rates, and drift buffer zone assessment [3, 18]. Pure droplet-tracking methods solve the Lagrangian equations of motion of discrete particle clouds with a stochastic component (random walk models), while treating the surrounding atmospheric wind field in a simplified, correlation-based manner [20, 21]. Full CFD studies add a further layer of complexity by solving the Navier-Stokes (NS) equations, usually additionally employing turbulence modelling for the wind field. They then take one of two primary approaches: Eulerian/Lagrangian (E/L) or Eulerian/Eulerian (E/E). The former computes the movement of a Lagrangian particle cloud using two-way interaction between the fluid and the discrete phase [24], whereas the latter computes droplet and wind phase as fluids via the NS, possibly utilizing interface capturing volume of fluid (VoF) methods [23].

Classes of problem with negligible diffusion are well-described by an E/L framework as numerical diffusive effects form an inherent part of E/E solution methods [25]. In particular, drag force-dominated application settings with constant interface topologies are preferentially formulated in the E/L view when compared to the "pure" two-fluid E/E approach. The phase fraction function employed in the classic two-fluid approach neglects any possible geometrical information about the phase boundary and treats each computational cell as a homogeneous mixture of both phases, which is not desirable in the description of a shape-sensitive force such as drag [26].

The individual tracking of large numbers of Lagrangian particles and their interactions as employed in E/L formulations is usually associated with higher computational cost than comparable E/E implementations; however, prior CFD studies show that the number of Lagrangian particles required to accurately model spray drift behaviour is relatively low [27]. Additionally, inclusion of evaporation effects does not significantly increase accuracy [28]. The volume fraction of the droplet phase ($\Phi_p$) encountered within agricultural sprays places the spray cloud in the upper region of dilute suspensions [29], providing justification for the dominance of E/L approaches in available spray drift CFD studies. The dominance of this approach is further explained by the easily accessible implementation of Lagrangian particle clouds in several CFD software packages, shipping with a wide variety of configurable subroutines. Considering the combined ease of use, physically justified range of applicability to the considered phenomenon and particularly the ability to directly specify several key properties of the droplet phase [30, 31], the E/L approach – in conjunction with experimental data for different nozzle types – was chosen to be the preferred framework for the formulation of the CFD setup in this study.

Experimental spray drift measurements have been carried out extensively. Nuyttens et al. carried out detailed measurement of spray drift variation with droplet size and nozzle type by use of spray collectors and phase Doppler particle analysis downwind of a single mounted nozzle in a wind tunnel [32]. Their wind tunnel setup follows ISO 22856:2008 [33] in line with governmental DRN accreditation schemes [10], and similar setups have been utilised in numerous experimental studies [27, 34, 35]. The extremely controlled nature of these wind tunnel experiments makes the data ideal for CFD validation. Further studies approach drift measurement through diverse methods, in-

Table 1: Existing spray deposition modelling techniques.

| | droplets | atmosphere | dynamic | time resolution | accuracy | commercial implementations |
|---|---|---|---|---|---|---|
| (1) Plume dispersion | concentration | matrix | no | months | low | ADMS |
| (2) Multiple regression | concentration | matrix | no | days | intermediate | DriftSIM |
| (3) Droplet tracking | discrete | correlated | yes | real time | high | AgDRIFT |
| (4) Euler/Lagrange CFD | discrete | Navier Stokes | yes | real time | very high | - |
| Euler/Lagrange SM | density | Navier Stokes | yes | real time | high (projected) | - |

cluding specialised test benches, other lab-based studies and measurements in the field [36–38]. The experimental setups of these studies vary significantly in terms of nozzle setups, atmospheric conditions and drift collection schemes, making comparsion to other studies complex.

The application of data-driven statistical surrogate modelling has been part of numerical fluid dynamics since the early days of the field. Kolmogorov, one of the founders of statistical learning theory and esteemed turbulence researcher, considered the turbulence phenomenon to be one of the most crucial application domains of statistical learning [39]. Large quantities of high-dimensional data were an everyday reality in the field long before the term Big Data was coined [39]. Applications of general Deep Neural Networks (DNN) [40], or more specific Recurrent Neural Networks such as Long-Short-Term-Memory (LSTM) [41, 42], Convolutional Neural Networks and Encoders (CNN) [42, 43] and Gaussian Process Regressions (GPR) [43] have shown promising results in the development of reduced order models. Most of these approaches however can be interpreted as image-to-image regression or image evolution models – modelling the evolution of some field quantity which can be interpreted as an image (such as the velocity field from synthetic CFD data) using an initial field as input parameter. Prior research results on whole image regression, or the prediction of an output image using only a few key physical parameters of the system as inputs, were found to be very sparse. While incredibly interesting for computational scientists, the development of accurate whole image regression models provides a difficult, yet exciting challenge – as will become clearer in the following pages.

# 3 Methods

## 3.1 CFD simulation of spray dispersion

The open source CFD toolbox OpenFOAM v6 was used to develop the computational implementation of the Eulerian/Lagrangian simulation of the spray drift case. Several tutorial cases (`AachenBomb, InjectionChannel, motorBike`) provided partial templates for the setup.

The layout of the wind tunnel used by Nuyttens et al. [32] was taken as a sensible basis for the domain of the CFD simulation, allowing comparability of results and validation of the numerical results. Domain geometry is indicated in figure 1. The standard `BlockMesh` mesh generation utility was used, with uniform grading in all three dimensions of the generated mesh. The final mesh resolution was set at $[70, 20, 20]$ resulting in 28,000 cells and 80,100 internal faces.

The discrete droplet phase was implemented as a Lagrangian particle model using the `kinematicCloud` class, which offers integrated support of a wide range of additional sub-models for particle forces, properties and injection models. In line with properties of standard pesticide formulations, particle properties were taken to be those of water at 25 °C and 1 atm. Drag forces were calculated using the `sphereDrag` sub method. Stochastic dispersion modelling via the `stochasticDispersionRAS` method, a stochastic `isotropyModel` and simple relaxation `dampingModel` completed the physical description of the Lagrangian particle cloud. A local `patchInteractionModel` was specified to allow escape of particles from the computational domain without any patch interaction.

The nozzle was consciously chosen not to be modelled using a specific mesh representation of a real nozzle – this would have also substantially increased the computational complexity of the simulation. A more generalised representation of the nozzle using the `InjectionModel` class provides a better basis to develop a more abstract model capable of simulating various nozzle makes, by specifying a set of characteristic input parameters obtained from experimental studies. The `ConeInjection` sub class implemented a disk-

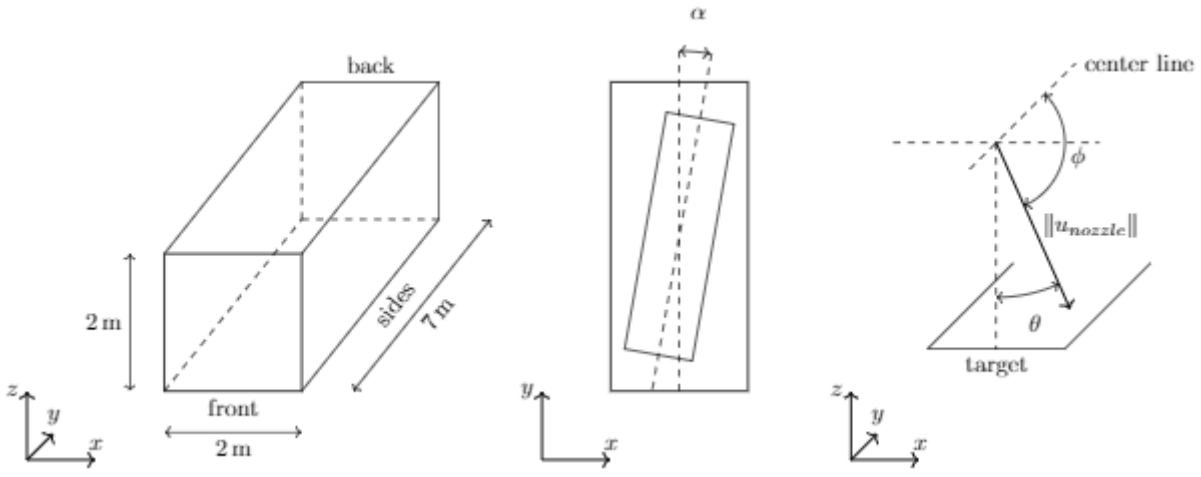


Figure 1: From left to right: computational domain; deposition target and wind angle; nozzle orientation

Table 2: Boundary- and Initial Conditions

| | $\boldsymbol{U}$ | $p$ | $k$ | $\epsilon$ | $\nu_t$ |
|---|---|---|---|---|---|
| ground | `noSlip` | `zeroGradient` | `kqRWallfunction` | `epsilonWallfunction` | `nutkWallFunction` |
| top | `fixedValue` | `zeroGradient` | `zeroGradient` | `zeroGradient` | `calculated` |
| sides | `fixedValue` | `zeroGradient` | `inletOutlet` | `inletOutlet` | `calculated` |
| front (inlet) | `fixedValue` | `zeroGradient` | `fixedValue` | `fixedValue` | `calculated` |
| back (outlet) | `inletOutlet` | `fixedValue` | `inletOutlet` | `inletOutlet` | `calculated` |
| internalField | $(\,u_x,\,0,\,0\,)\,\mathrm{m\,s^{-1}}$ | $0\,\mathrm{Pa}$ | $1.3\,\mathrm{m^2\,s^{-2}}$ | $1\,\mathrm{m^2\,s^{-3}}$ | $0\,\mathrm{m^2\,s^{-1}}$ |

shaped source of 10 mm diameter, emitting $5.7 \times 10^4$ particles of the `kinematicCloud` at a parcel size of one over a spray angle of 80°. The nozzle outlet was located 1 m above the floor of the domain and 0.5 m from its front. Parcels were specified on the basis of mass, with the density being fixed as described above and the droplet diameters being distributed according to a Rosin–Rammler distribution. Hence the variable input parameters for the nozzle are given by particle diameter distribution's parameters, the initial particle velocity and the nozzle orientation via the azimuthal and polar angles.

Turbulence was modelled using a $k$–$\epsilon$ model, with initial turbulent kinetic energy and dissipation rate values being based on atmospheric measurements [44]. $k$–$\epsilon$ is known to under-predict effects of turbulence on lateral dispersion [31], which was aimed to be compensated by introduction of a stochastic component with `StochasticDispersionRAS`. Boundary and initial conditions were set as outlined in table 2 under the assumption that the domain is wide enough to employ far-field boundary conditions on the top and side faces of the domain. The initial velocity field and the fixed boundary values are given by a unidirectional velocity field of magnitude $u_x$, corresponding to the mean wind speed. The case was then solved with the incompressible, two-way interaction solver `MPPICFoam` using the divergence schemes Gauss upwind for $k$ and $\epsilon$, Gauss linear upwind for $\boldsymbol{u}$ and Gauss linear corrected for the Laplacian scheme.

The ground deposition density data was extracted from the Lagrangian particle trajectories using the `ParticleCollector` function object, which allows the user to specify a polygonal surface and "collect" the total mass of particles having passed through it. A self-developed Python utility function was used to generate a rectangular grid of specified resolution of independent `ParticleCollector` instances. A "deposition target" resembling a row of crop of size 6 m by 2 m was defined upon which the `ParticleCollector` grid was placed. To simulate non-frontal winds without altering the CFD setup, the nozzle, the target and correspondingly the `ParticleCollector` grid were rotated around the target's center by an angle $\alpha$, providing a lateral wind field from the perspective of the target / row of crop.

The data was then extracted from the individual instances using a further Python utility function outputting a single `NumPy` matrix of dimension $[x_{\text{cells}}, y_{\text{cells}}, t_{\text{cells}}]$, with $t$ being the number of time step outputs – equivalent to a collection of $t$ snapshots of the deposition density field at a grid resolution of $[0, x_{\text{cells}}] \times [0, y_{\text{cells}}]$. These snapshots can be interpreted as images of the deposition field, with each cell's spray deposition value corresponding to a pixel's colour scale of the image, and will be referred to as as such throughout.

Mesh independence studies were performed on the simulation to ensure adequate mesh density. The main output variable of interest was the deposition density distribution – therefore, the variable of interest for comparison of mesh densities is this density distribution. The number of mesh elements in a simulation was varied, keeping element aspect ratio identical. These simulations were then run and variation in the variable of interest (drift percentages) was then calculated from `ParticleCollector` and average mean square error between each between this output for each simulation against the next least dense simulation calculated. The final timestep is taken to indicate steady state. Drift percentages ($D\%$) were calculated as the ratio of mass deposited ($m_d(x_i)$) on cell $x_i$ to total mass introduced ($m_{total}$):

$$D\% = \frac{m_d(x_i)}{m_{total}} \times 100 \tag{1}$$

Validation of CFD results against experimental data was performed using 3 independent data sets. Spray cloud properties for three individual nozzles were taken from a detailed phase doppler particle analysis study by Nuyttens et al. [32] and used to derive the shape and scale parameters of the underlying Rosin–Rammler Distribution. The geometric dimensions of the validation simulations were in accordance with ISO22856 [45] and the setups of the experimental studies. A `ParticleCollector` grid with made 0.1 m sided squares parallel to the frontal wind was placed at a height of 0.1 m, corresponding to usual foliage height.

## 3.2 Computational pipeline

The validated CFD model was then integrated into a `Nextflow` (a workflow manager for the creation and

execution of data-driven computational pipelines in a fluent domain-specific language) data pipeline. This enabled a controlled workflow for the computation of over 5500 simulations of automatically varied parameters which can be specified in a single file. Additionally, `Nextflow` allowed native batching of runs on the Imperial College High Performance Cluster, while also enabling parallel computing at the process level.

A `Nextflow` channel iteratively executes a process for every set of parameters in an input `.csv` file. The process itself is a `bash` script which is executed in a dynamically created work directory, into which the OpenFOAM template case and the utility scripts are copied. Several self-developed Python scripts, making use of the `pyFoam` module, are called automatically to read the run parameters, create the specific OpenFOAM case and the `ParticleCollector` instances. The case is run using `MPPICFoam` and further Python scripts are called to extract a `.npy` file of particle collector outputs over time. A minimum working example of the final computational pipeline is provided in a GitHub repository [46].

Table 3 lists the constraints imposed on the five input variables specifying wind velocity and direction, nozzle orientation, and exit velocity (with the particle size distribution being fixed for a given nozzle). Samples were then drawn from the resulting space via Latin hypercube sampling, using a space filling sampling criterion, and processed via the computational pipeline.

Table 3: Ranges of simulation input sample space

| | lower | upper | unit |
|---|---|---|---|
| wind velocity | 0.5 | 6.5 | $\mathrm{m\,s^{-1}}$ |
| wind angle | 0 | 10 | ° |
| nozzle $\theta$ | -45 | 45 | ° |
| nozzle $\phi$ | -45 | 45 | ° |
| exit velocity | 2.5 | 3.5 | $\mathrm{m\,s^{-1}}$ |

## 3.3 Surrogate model construction

### 3.3.1 Black-box approaches

Initial exploratory surrogate model construction was performed using a variety of regression models from the `scikit-learn` module. Input parameters were scaled using the `StandardScaler` procedure, and the individual $[30, 20]$ matrices were flattened to a drift-percentage normalised one-dimensional output vector of length 600. After randomly splitting the dataset in a train and validation set (80/20), several Multilayer Perceptrons (MLP), Random Forest Regressors (RFR) and a Gaussian Process Regressors (GPR) models were trained using different hyperparameter variations – each attempting to predict the entire image from the five input parameters.

In addition, pixel to pixel (P2P) ensemble regression was performed by training an individual MLP and GPR for each pixel of the particle collector, as opposed to training one model for the whole collector. This was designed to reduce the effect of information density issues and complexity of whole-image output scales (see 5.2.1). The lack of success of these approaches lead to development of more advanced surrogate modelling methods, as outlined in the following paragraphs.

### 3.3.2 Pesticide drift surrogate model

A custom image regression algorithm was formulated for surrogate modelling of pesticide drift. A pixel frame of size 2x2 traverses the image at a stride of (2,2). In each frame, a local coordinate system $\{0,1\} \times \{0,1\}$ is defined and treated as an additional pixel feature, providing relative spatial positioning information of the pixels to the model. Smoothness is imposed by including the current mean pixel value, and assigning it a relative position at the centre of the frame $(0.5, 0.5)$. The input vector is then formed by concatenating the vector of run parameters with the local pixel coordinates for the five points, while the output vector is constructed by stacking the vectors of pixel values and the computed mean vector. A GPR was then trained for each frame, fundamentally providing the model with built-in prediction uncertainty quantification. At the specified resolution, this results in training 150 individual frame models – each only focusing on a subset of the image, but from which the whole image can be easily reconstructed. Regression scoring on the test set is conducted on a frame by frame basis, resulting in an image of coefficients of determination providing more holistic information of true regression performance.

Conceptual unit testing of the proposed algorithm was conducted on synthetically constructed dispersion fields to provide a benchmark on less chaotic data. The data was generated from Gaussian' curves with a constant mean and standard deviation following geometric Brownian motion with positive drift in wind direction. White noise was added to all data points, and the proposed algorithm, a P2P GPR and whole-image GPR were trained as benchmarks. The proposed algorithm was then applied to a CFD data set with a compressed parameter space (nozzle angles were restricted to $\pm\,10°$ and wind velocity only up to $5\,\mathrm{m\,s^{-1}}$) to produce a less chaotic dataset and evaluated for its performance.

### 3.3.3 Disinfectant drift surrogate model

The surrogate modelling task was reformulated for disinfection application as a binary decision of sufficient application of disinfectant volume at some point on the spray target. The CFD output images were processed into binary images using a defined threshold of $7.9\,\mathrm{ml\,m^{-2}}$, corresponding to both a deposition rate higher than one obtained by uniform deposition across the entire target, and slightly above the recommended application volume of spray mist surface disinfection with dilute hydrogen peroxide. The pixel by pixel approach was to predict a binary variable for

each pixel from the 5 input parameters – again training 600 individual classifiers on the entire thresholded CFD dataset, pixel to pixel. A neural network with two hidden layers at 25 neurons each, using a hyperbolic tangent as activation function and lowered convergence tolerance of $10^{-6}$, was chosen as the final model architecture. Performance was evaluated via the cross-image mean accuracy, the position dependent accuracy as an accuracy image, and the false positive rate – particularly crucial in the disinfection application.

# 4 Results

## 4.1 Validation of the CFD simulation

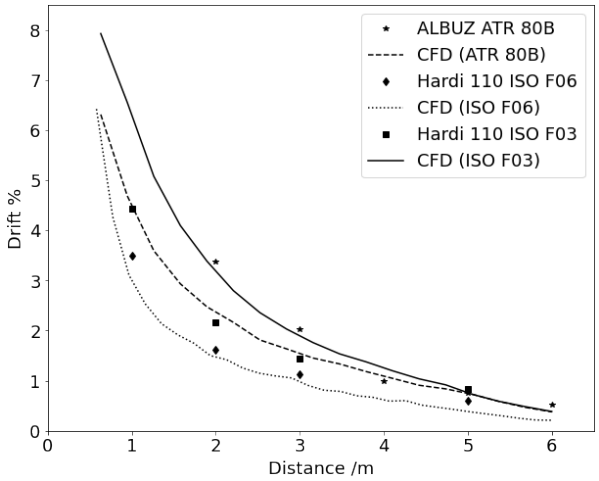


Figure 2: Validation curves for three nozzle designs

The CFD simulation required both validation against experimental results from literature, and verification of mesh independence. Validation results are presented in figure 2. CFD-predicted drift percentages at varying distance downwind from the nozzle (lines) are compared to experimental results for three different nozzle types (points) from three literature sources [27, 34, 35], showing good adherence of the model's predictions to all three nozzle types. Mesh independence analysis showed a mean square error (MSE) of the resulting deposition fields between successive mesh gradings of $[63,18,18]$ and $[70,20,20]$ to be $<1.8\times10^{-5}$; hence the mesh grading of $[70,20,20]$ with $2.8\times10^{4}$ elements was deemed sufficient.

## 4.2 Surrogate modelling

### 4.2.1 Exploratory analysis of the generated training data

In total, data from more than 5500 individual CFD simulations was successfully generated – amounting to more than 100.000 snapshots and $6\times10^{7}$ data points. The resulting deposition fields showed an astonishing variety of spatial structures.

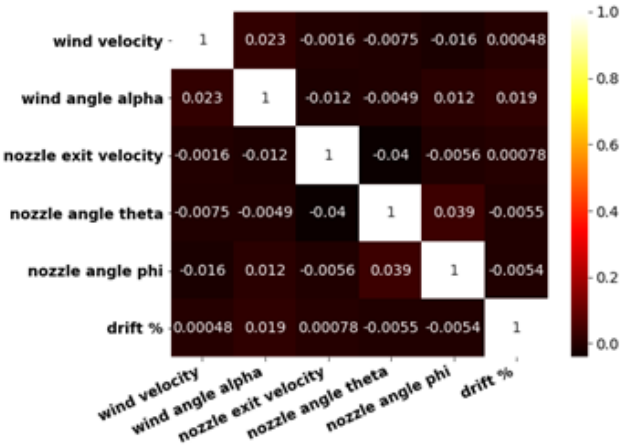


Figure 3: Correlation matrix for entire training dataset – note near-zero correlations for all variables

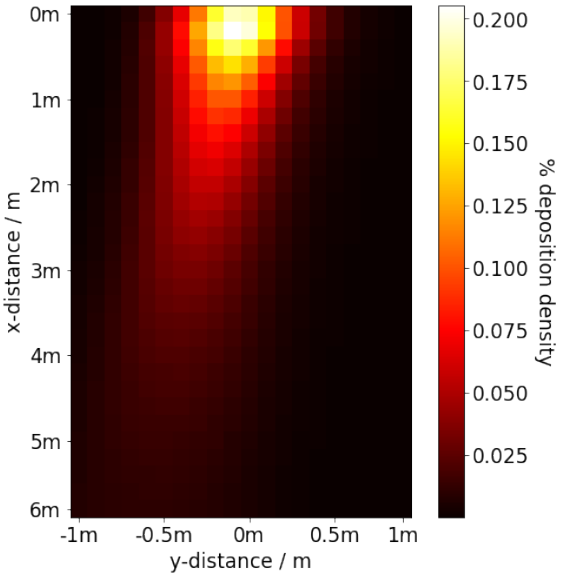


Figure 4: The mean deposition density image of all generated 5500 CFD simulations – note skew of mean to left due to variation of wind angle only in this direction relative to head-on

Correlations between individual cell deposition densities and input parameters (figure 3) were very low, further illustrating and quantifying the complexity of the underlying phenomenon. The large variety of plume shapes seems promising with regards to application of active control to influence downwind spray deposition, with the nozzle orientation in particular having a very intricate interaction with the resulting deposition field and hence ability to counteract wind influences – however, it also provides testimony to the challenging problem of modelling these interactions in a surrogate model. The mean image of all generated snapshots can be found in figure 4. The mean approximately follows simple Gaussian plume dispersion behaviour on the first metre, but deviates significantly further downwind. It is interesting to note that while

the mean at first follows the expected diagonal line along the centre of the wind angle variation, deposition behaviour further downwind becomes increasingly dispersed – from visual inspection of several deposition density images, it seems that the effects of the nozzle orientation and wind direction interplay are particularly pronounced in the more dilute region of the spray.

#### 4.2.2 Black box approaches

Black box whole-image regression techniques showed poor results. Reduction of position dependent accuracy metrics to a single scalar by aggregation functions such as the mean $R^2$ or root relative mean square error (RRMSE) indicated low performance off all 3 regressors (MLP: $R^2 < 0$, RRMSE 20%; RFR: $R^2 < 0$, RRMSE 11%; GPR: $R^2 < 0$, RRMSE 29%), but pre-aggregation performance data showed significant differences in areas of low vs. high deposition(RRMSE $< 0.1\%$ vs. $> 40\%$).

Classic regression metrics showed higher performance of the P2P approach over the whole-image model: the average prediction error was reduced (MLP RRMSE $< 18\%$). However, visual inspection of the P2P model revealed no recognition of the spatial image structures and characteristic plume shapes. While this clearly showed the need for the more advanced surrogate modelling techniques, it also further illustrated the difficulty of formulating appropriate loss functions for position-dependent output accuracy, providing a challenge both in assessment of the model and appropriate application-specific formulation of the underlying optimisation problem of the machine learning algorithms.

#### 4.2.3 Pesticide drift surrogate model

After promising performance in the proof-of-concept test ($R^2$ proposed 0.34 vs P2P, whole-image $\approx 0$), the proposed algorithm was applied to the described CFD data set of 400 runs with a reduced parameter space. The RRMSE image, shown in figure 5, shows particularly strong performance of the regression in the the key downwind area of the deposition density image. The maximum RRMSE across the image on the test split of the reduced data set was 1.3%, ensuring good accuracy throughout and placing the model in a range appropriate for active control purposes. The $R^2$ score indicates that more than 80% of the variance found in the data set can be accounted for by the model. Given the very low RRMSE, the lower-than-expected $R^2$ value might be a result of the imposed smoothness condition introducing additional variance, with the calculated mean being considered a frame feature. This possibly indicates a smoothness/variance trade off – partly caused by the discrete, comparatively coarse grid nature of the `ParticleCollector`. The proposed algorithm compares favourable to the whole-image and the pixel to pixel regression attempts, with whole image GPR regression on the reduced parameter set performing more than four times worse than the proposed algorithm (but significantly better than on the large dataset; RRMSE 8%) and pixel to pixel regression also performing better on the reduced parameter dataset, but also significantly worse than the proposed algorithm with an achieved RRMSE of 11%. Visual inspection of CFD ground-truth output / surrogate output pairs also confirmed the better performance of the proposed algorithm in correctly recognising and reproducing physically sensible plume shapes, of which a particular example is also illustrated in figure 5.

#### 4.2.4 Disinfectant drift surrogate model

The classification model for spray drift disinfection achieves an overall average accuracy of 78% with a false positive rate of 12%, at a standard decision threshold of 0.5. The accuracy image (figure 6) shows an interesting spatial trend with the area of lowest accuracy being present in the dilute border region of the spray, and ac-

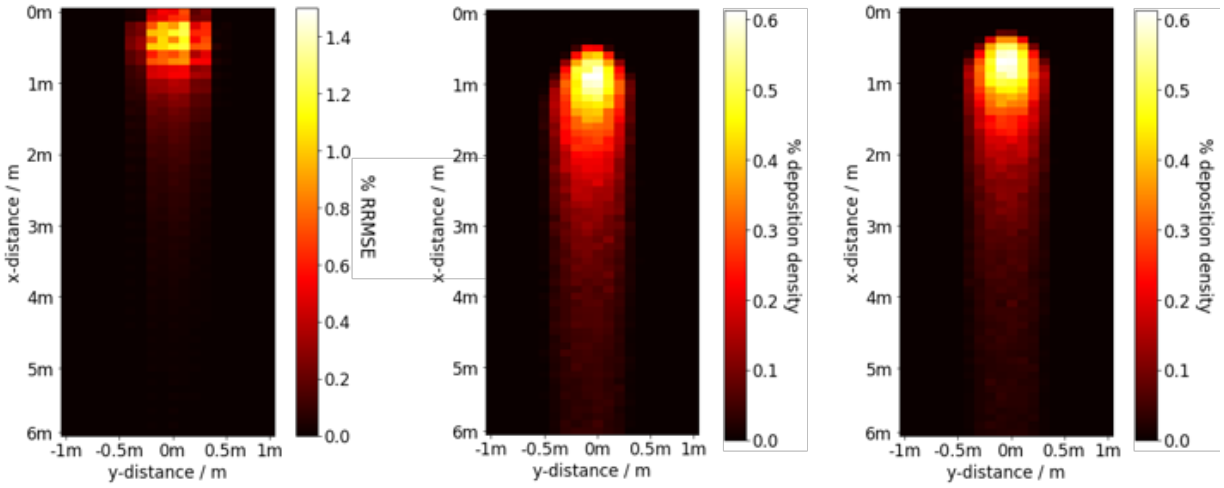


Figure 5: From left to right: relative root mean square error (RRMSE) of the proposed whole image regression algorithm; an exemplary CFD deposition density output image compared to; the deposition density image predicted by the proposed algorithm

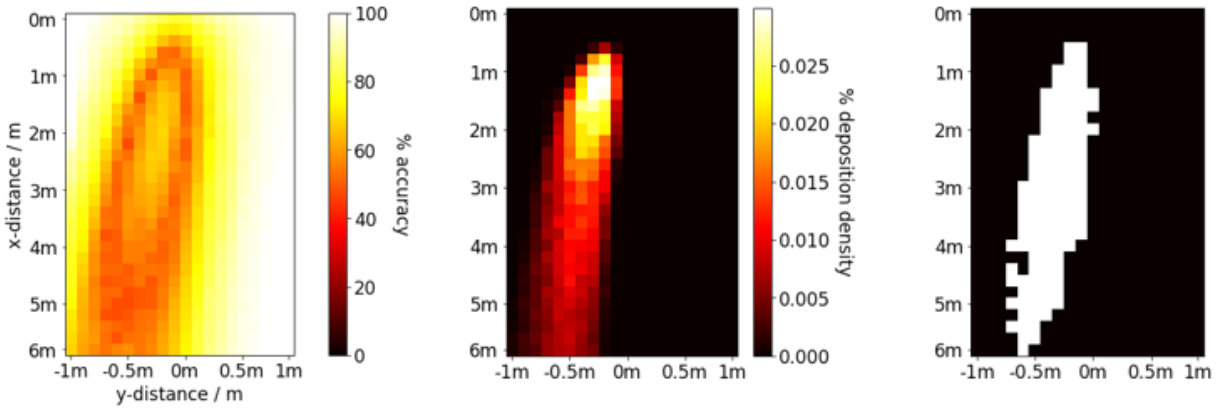


Figure 6: From left to right: classifier accuracy of the proposed disinfection threshold classification algorithm; an exemplary CFD deposition density output image compared to; the binary disinfection image predicted by the proposed algorithm with white indicating areas of sufficient disinfectant application

curacy increasing both inwards towards the plume and the outermost spray regions. The current threshold, set above the recommended application volume limit, mitigates the current false positive rate by providing a safety margin. The classifier can be easily adjusted to incorporate different safety margins and or decision thresholds, and could in it's current state be configured to provide a meaningful active control model for UAV spray mist surface disinfection.

# 5 Discussion

## 5.1 Efficacy of CFD models

Comparison to literature data indicated that the CFD model is accurate to those scenarios tested. Easy transferability of the `kinematicCloud` and `coneInjection` modelling approach to different nozzle makes was indicated by the testing for different nozzle designs. However, experimental data for every single permutation of input parameter is not available – it can be predicted that, as physical variations compared to these cases were relatively limited (e.g. Reynolds numbers remain in similar ranges), the validity over the tested sample space is acceptable, but this could not be absolutely verified. Additionally, the model was formulated for comparability to literature studies as opposed to specific industrial scenarios (e.g. 20 nozzles on a tractor-towed pesticide sprayer), although the method could certainly be adapted to this by future researchers.

## 5.2 Surrogate modelling

### 5.2.1 Model formulation issues

Initial whole-image MLP and GPR showed a stagnant decrease of training loss, relatively non-smooth predictions and their predicted output images converged to the mean of all training images, usually indicating that the model is not sufficiently complex to capture the underlying relationship. RFRs produced output features with more noticeable spatial distinctions, but also did not capture the true image structure sufficiently to be chosen as the most applicable algorithm.

Additionally, the commonly used regression metrics and loss functions showed limited applicability to whole image regression. $R^2$ values in image regression would become position-dependent, and hence effectively also give an 'image of determination'. $R^2$ and RRMSE figures for these approaches indicated poor overall performance, hence development of the other approaches to tackle the problem. Three key issues leading to low regressor performance were identified:

1. Low information density for prediction of 600 values from just five input parameters
2. Extremely low covariances between the feature space and the deposition data
3. Large variations in magnitude of output values, which cannot be uniformly scaled

P2P ensemble regression was used as a simple potential solution to these difficulties. This was effectively a failure as it did not recognise the spacial structure of the images. This is obscured by removing the spatial relationships between pixels and the effects of noisy variation in pixel values are even more apparent. It became clear that this type of application requires a specifically constructed approach to surrogate modelling. The main challenges to whole-image regression in this application setting are thus summarised:

1. Imposition of output smoothness and compactness
2. Imposition of spatial structure information
3. Robustness of the method to scale of parameters
4. A position-dependent goodness-of-fit assessment metric

Revisiting the requirements of the two use cases of a spray drift surrogate model outlined in the introduction revealed some differences. For disinfection and decontamination procedures, the key performance measure can be interpreted as a binary decision: whether a

sufficient amount of active component has been applied to a specific area. For drift-reducing active control of agricultural sprays however, a detailed pesticide deposition field output is required. A individually tailored ML approach to each of these differing requirements was deemed necessary to achieve the required accuracy. A disinfection spray drift surrogate model can hence be formulated as a classification problem, using some deposition threshold of applied active component per unit area; the classification approach would allow working on a simple $\{0,1\}$ output scale with reduced importance of the wider spatial structure, while having clearly defined performance metrics in the accuracy and false-positive rate. The pesticide drift surrogate must be a regression model that can – at least in some localised sub-domain – learn from the spatial structures and provide a smooth output result, while preferably being capable of built in uncertainty quantification.

#### 5.2.2 Pesticide drift surrogate model

The traversing local frame regression showed highly promising results in the current application as well as general whole image regression; the ability to preserve the spatial relationships within the image whilst reducing the effects of a lack of information density and local output scale comparability. Regression scoring showed strong results on both synthetic Gaussian data and CFD data, providing a near instantaneously evaluable surrogate model of adequate accuracy for active control application – albeit on a reduced physical parameter space. Further research is required to extend the proposed algorithm to larger, more complex CFD datasets. The primary next step would be the development of a simple to use, modular code base for the proposed algorithm enabling investigation into different hyperparameter combinations, frame sizes and strides, as well as possible whole image regression scenarios outside the spray drift phenomenon.

#### 5.2.3 Disinfection drift surrogate model

The constant threshold value is a key explanation behind the shape of accuracy plot (figure 6), the value of $7.9\,\mathrm{ml\,m^{-2}}$ having been chosen as a good fit overall for most images. ROC-AUC optimisation of decision thresholds for individual classifiers should enable further decrease of the false positive rate, by reflecting the preference of specificity over sensitivity for a disinfection application. This would increase descriptiveness of plume shape, but complicate the training methodology due to the ensemble nature of the classifier and associated worries of local overfitting. This technique was overall shown to be a strong predictor of both plume shape and the key performance indicator of the desired application setting. The surrogate model hence produces good estimates of surface coverage in only a few miliseconds, compared to the average computing time of two hours of the developed CFD simulation. Further research would likely provide potential for additional accuracy improvements focusing not only on ROC tailoring but also Bayesian hyperparameter optimisation of the underlying neural network.

#### 5.2.4 Numerical assessment of descriptiveness of shape

Although assessment of shape sensitivity of models was made visually, a numerical method for comparison of this was not found. The hence not optimal loss function used in the algorithmic training of ML methods impacts performance as networks better describing the true shape cannot be directly searched for. It is assumed that increased $R^2$ and reduced RRMSE will in general indicate better shape descriptions: this may not always be true, however. For example, a model could have significant variation in every pixel compared to the true values, resulting in poor RRMSE values, but describe overall plume shape relatively well. If a metric of description of shape could be formulated as a loss function, model training and selection could be significantly enhanced and help find drift models of lower generalisation error in the complex parameter space. This would perhaps be more effective when training on larger images; converting the 600 pixel images used here to edge vectors, the first step in numerical shape description, was found not be effective due to the coarse variation in spatial structure.

## 6 Conclusion

In the course of this investigation, a validated CFD model for spray drift from a single nozzle was produced, a custom computational pipeline for generating CFD data on remote clusters was produced and various approaches to generating surrogate models of spray dispersion were evaluated. In this case, black box models used for whole-image regression showed no efficacy in predicting spray patterns and pixel-to-pixel regression on an unprocessed image failed to predict spatial structures within the image. The traversing local frame regression approach described spatial structure and was overall an effective predictor of spray behaviour. Similarly, the thresholding and classification-based approach described plume shape well. Bayesian hyperparameter optimisation could be used to improve on these results. Improved implementation of this modelling strategy could show utility in the active control of spray dispersion.

More radically, the problem could be rephrased for image-to-image regression, with an additional particle collector placed under the nozzle, perhaps an image relayed from a camera on a real-life nozzle. This would help solve the information density problem faced by the surrogate modelling approaches and potentially reduce the sensor array needed in a real-life control problem.

The computational tools developed for this project are publicly available for adaptation and use in further research [46].

# 7 Acknowledgements

Special thanks is given to Dr Lachlan Mason, Gabriel Farah, Dr Indranil Pan and Dr Themistoklis Botsas for their invaluable aid and support with this project.

# References

(1) Sharma, A. et al. *SN Applied Sciences* **2019**.
(2) Pimentel, D.; Burgess, M. In *Integrated Pest Management*, 2014.
(3) Felsot, A. S. et al. *Journal of Environmental Science and Health - Part B Pesticides, Food Contaminants, and Agricultural Wastes* **2011**.
(4) Stephenson, G. R. et al. *Pure and Applied Chemistry* **2006**.
(5) Nuyttens, D. et al. *Communications in agricultural and applied biological sciences* **2009**.
(6) Hilz, E.; Vermeer, A. W. *Crop Protection* **2013**, *44*, 75–83.
(7) De Snoo, G. R.; De Wit, P. J. In *Ecotoxicology and Environmental Safety*, 1998.
(8) Matthews, G. *Outlooks on Pest Management* **2008**, *19*, 278–279.
(9) Van de Zande, J. et al. *Spray drift and off-field evaluation of agrochemicals in the Netherlands*; tech. rep.; Plant Research International, 2007.
(10) Rautmann, D. In, 2013.
(11) Stafford, J. V. *Journal of Agricultural and Engineering Research* **2000**.
(12) Walter, A. et al. Smart farming is key to developing sustainable agriculture, 2017.
(13) Mahbub, M. *Internet of Things* **2020**.
(14) Lottes, P. et al. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp 3024–3031.
(15) Skotadis, E. et al. *Computers and Electronics in Agriculture* **2020**, *178*, 105759.
(16) Eddy, Y. et al. *Revista Argentina de Clínica Psicológica* **2020**, *29*, 221.
(17) EagleHawk Drone-Enabled Disinfectant Spraying `https://www.eaglehawk.io/drone-enabled-disinfectant-spraying` (accessed 12/21/2020).
(18) Bache, D. H.; Sayer, W. J. *Agricultural Meteorology* **1975**.
(19) Zhu, H. et al. *International Conference on Pesticide Application for Drift Management* **2004**.
(20) Teske, M. E. et al. *Environmental Toxicology and Chemistry* **2002**.
(21) Butler Ellis, M.; Miller, P. *Biosystems Engineering* **2010**, *107*, 169–177.
(22) Duga, A. T. et al. *Biosystems Engineering* **2017**, *154*, Engineering Approaches for Reducing Spray Drift, 62–75.
(23) Parra, H. G. et al. In *New Knowledge in Information Systems and Technologies*, ed. by Rocha, Á. et al., Springer International Publishing: Cham, 2019, pp 812–822.
(24) Zhang, B. et al. *Biosystems Engineering* **2018**, *166*, 184–199.
(25) Subramaniam, S. Lagrangian-Eulerian methods for multiphase flows, 2013.
(26) Cristini, V.; Tan, Y. C. Theory and numerical simulation of droplet dynamics in complex flows - A review, 2004.
(27) Baetens, K. et al. *Computers and Electronics in Agriculture* **2007**.
(28) Woods, N. et al. *Journal of environmental quality* **2001**, *30*, 697–701.
(29) Elghobashi, S. *Applied Scientific Research* **1994**, *52*, 309–329.
(30) Arshad, S. Report - Descriptions and comparisons of sprayFoam, reactingParcelFoam, and basicSprayCloud, basicReactingCloud, 2014.
(31) Nygren, A. Modeling of Spray Formation and Development in OpenFOAM with Application to Diesel and Alcohol Fuels, 2018.
(32) Nuyttens, D. et al. *Biosystems Engineering* **2009**.
(33) BS ISO 22856:2008: Equipment for crop protection. Methods for the laboratory measurement of spray drift. Wind tunnels, eng, 2009.
(34) Nuyttens, D. et al. *Communications in agricultural and applied biological sciences* **2008**.
(35) Torrent, X. et al. *Crop Protection* **2017**.
(36) Nuyttens, D. et al. *Biosystems Engineering* **2017**.
(37) Nuyttens, D. et al. *Transactions of the ASABE* **2009**.
(38) Butler Ellis, M. C. et al. *Biosystems Engineering* **2010**.
(39) Brunton, S. L. et al. *Theoretical and Computational Fluid Dynamics* **2020**, *34*, 333–337.
(40) Kutz, J. N. *Journal of Fluid Mechanics* **2017**, *814*, 1–4.
(41) Mohan, A. T.; Gaitonde, D. V. A Deep Learning based Approach to Reduced Order Modeling for Turbulent Flow Control using LSTM Neural Networks, 2018.
(42) Hasegawa, K. et al. In *ASME-JSME-KSME 2019 8th Joint Fluids Engineering Conference, AJKFluids 2019*, 2019.
(43) Maulik, R. et al. Latent-space time evolution of non-intrusive reduced-order models using Gaussian process emulation, 2020.
(44) Han, J. et al. An Estimation of Turbulent Kinetic Energy and Energy Dissipation Rate Based on Atmospheric Boundary Layer Similarity Theory, 2000.
(45) BS ISO 22866:2005: Equipment for crop protection. Methods for field measurement of spray drift, eng, 2006.
(46) Klotz, D.; Dale, S. driftml GitHub repo `https://github.com/klotzd/driftml`.