Appendix 1
Approximating Browse exposure outputs

Marc Kennedy, 30-04-13

1. Introduction

This report describes a fast approximation to the Silsoe spray drift model created as part of the Browse project. The main aim of Browse is the creation of a flexible and comprehensive software tool for estimating multiple exposure types under a variety of scenarios, which can also account for uncertainty and variability in input parameters. In order to achieve this, it is essential to produce a fast model that can be run many thousands of times for any user-specified conditions selected from a wide range. The approximation errors should be as small as possible, but this requirement must be balanced against the need for practical implementation and fast calculations during real-time use.

Outputs from the Silsoe model are required to model exposures to bystanders, residents and opertors but due to its complexity it is not practical for use directly in the Browse tool. In the next section we describe a model to approximate these outputs, so that an output value can be estimated for any required input parameter set.

2. Model description

The focus of our analysis is the spray drift model of Butler Ellis and Miller (2010), adapted from the earlier model of Miller and Hadfield (1989). Five different types of output can be generated: Ground deposit (ml m^{-2}), Adult inhalation factor (ml m^{-3} s), Child inhalation factor (ml m^{-3} s), Adult airborne spray (ml m^{-2}) and Child airborne spray (ml m^{-2}). Henceforth these will be labelled Ground, IfC, IfA, AirC, and AirA, respectively. Details of the model are given in Butler Ellis and Miller (2010) and Kennedy *et al* (2012). The results shown here relate to a single nozzle type but the same process can be repeated for others.

2.1 Data for statistical analysis: Inputs and outputs

Each pairing of the 5 spray drift outputs types was with 4 crop heights (0.1, 0.5, 1.0, 1.5 m) was considered, giving a total of 20 outputs to estimate. Preliminary analyses and previous studies within the Bream project suggested that these outputs should be estimated independently, as the input-output relationships can be qualitatively different. Sufficient training data can be obtained to enable individual estimates to be derived, particularly as the model generates multiple output types per run.

The same 5 inputs can be varied in each of our approximate models (number of nozzles, boom height, forward speed, wind speed and wind angle). The parameters are listed in Table 1, with ranges selected to include reasonable bounds for potential future user choices. For each of the 20 output types, 600 sets of training inputs were generated from within the 5-parameter input space. The first 400 input sets were dispersed within the full input ranges, so that prediction is possible beyond the range of 'typical' conditions. A further design was generated using narrower input

ranges, to give more accurate approximation in those regions of the input space likely to be most common. The computer program GEM-SA (Kennedy, 2005) was used to generate the 2 maximin latin hypercube input designs (Morris and Mitchell, 1995). Figure 1 below shows the marginal pattern of points for each pair of inputs. The original code was run for each input set, to generate 600 corresponding curves, representing the output at distances 1-20 m. Some examples are plotted in figures 2, 6, 7, 8.

Some initial cleaning of the data was carried out. First, a spline smoother was applied to each curve to remove the effect of stochastic 'noise' in the calculations. In addition, those runs which include one or more zero values in the output were removed from the dataset. The latter was to prevent numerical problems when analysing log-transformed outputs.

It was considered important to estimate the entire output (as a function of distance) for a single input configuration, as these need to be presented as a continuous curve if outputs are require at multiple distances. Estimation of exposure curves was considered in two parts. The first part is the 'source term', which is the output at the single distance (2m). The second part is the exposure curve relative to the source term as a function of distance. The accuracy of the initial source term is critical for overall accuracy, and the relative decay curves were found to have similar forms. The models for both parts are described in the following subsections.

2.2 Emulation of source term

Bayesian emulation was applied to the output close to source (2m). This model provides flexibility to adapt to complex relationships between the inputs and outputs as it is not constrained to a particular form e.g. linear or quadratic. It is designed to estimate a single output that is a smooth function of the model inputs. The GEM-SA software was used to generate the user inputs for the training data and also to analysis the results, allowing us to predict log(output at 2m, input = $\boldsymbol{x}$) for any value of $\boldsymbol{x}$ within the ranges specified for each of its input component. GEM-SA is restricted to 400 training inputs so we combined the first 200 points from the full input range set with the 200 from the more focussed input range set. A small numerical error or nugget term was included in the GEM-SA analysis to improve the fit. The output is known to have some stochastic variation due to random processes, meaning that repeated runs of the model at the same input point can give very slightly different outputs. The use of the nugget term smooths this out and produces a more accurate estimate of the underlying mean response. We also see predictions that are smoother than the code outputs, as seen for example in figures 6, 7, 8.

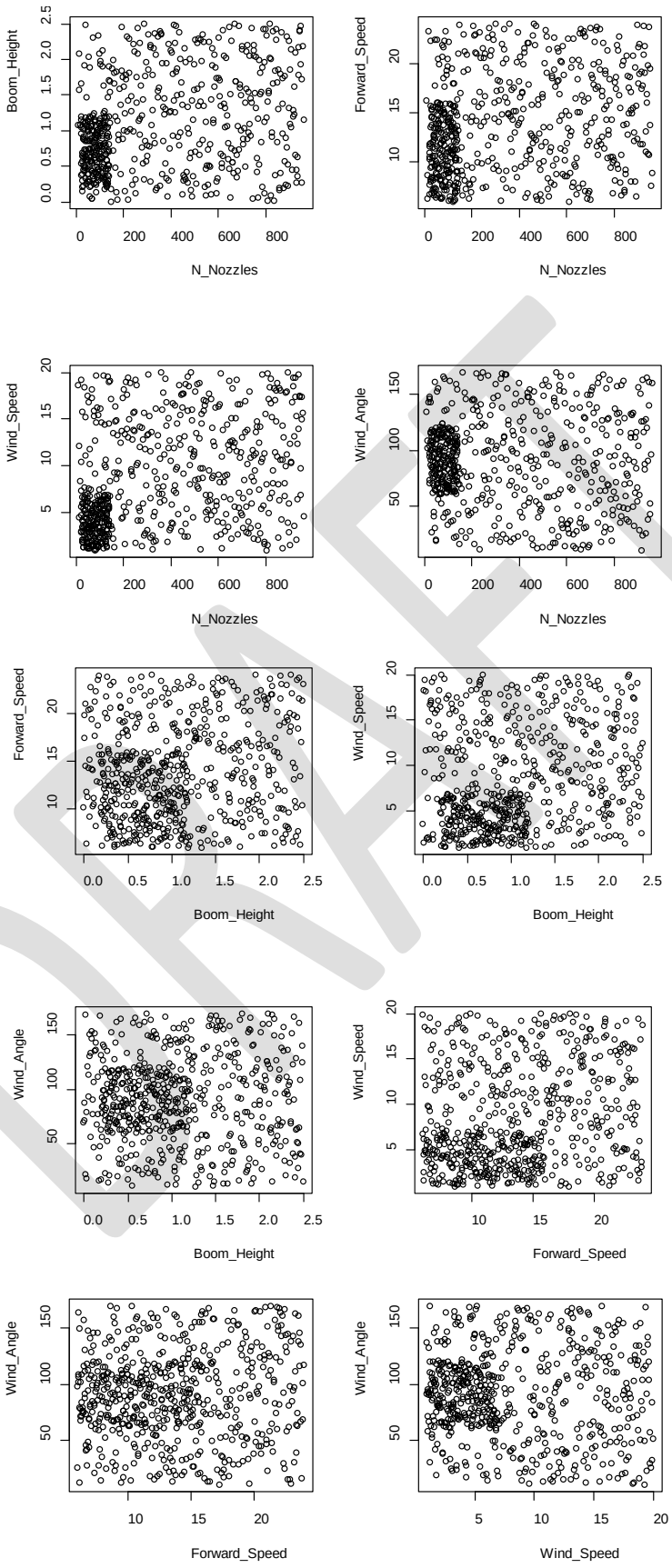
Boom_Height
N_Nozzles
Forward_Speed
N_Nozzles
Wind_Speed
N_Nozzles
Wind_Angle
N_Nozzles
Forward_Speed
Boom_Height
Wind_Speed
Boom_Height
Wind_Angle
Boom_Height
Wind_Speed
Forward_Speed
Wind_Angle
Forward_Speed
Wind_Angle
Wind_Speed

Figure 1: Input designs used in the emulator

2.3 Model for decay curve

A generalised linear model was used to characterise the decay as a function of distance and other variables (Table 1), relative to the source term estimated in Step 1. On inspecting the output curves it was evident that there were distinct classes of functions amongst the training runs. Most were decreasing as a function of distance, but others were increasing at short distances before gradually levelling off and decreasing.

We experimented with various forms, including transformations on input/outputs and terms in the regression equation. Cross-validation, Quantile-Quantile and residual plot diagnostics were used to assess and iteratively improve the model fits. We assume distance > 2m and make the transformation d = ln(distance) – ln(2) and used log(windSpeed).

Three generalised linear models were used to estimate the ratio $r(\boldsymbol{x}, d) = f(\boldsymbol{x}, d)/f(\boldsymbol{x}, 0)$. Note that $r(\boldsymbol{x}, d) \to 1$ as $d \to 0$ so the models must respect this condition.

1. The probability π_i of an increasing curve for general input x_i is modelled as

$$logit(\pi_i) = \boldsymbol{h}_c(\boldsymbol{x}_i, d_i)^T \boldsymbol{\beta}_c + \varepsilon_i^c$$

where we use the transformed logit scale $logit(\pi) = \log(\pi/(1-\pi))$, $\boldsymbol{h}_c()$ is a vector of relevant regression terms depending on the model inputs, $\varepsilon_i^c \sim N(0, \sigma_c^2)$ are independent error terms and $(\boldsymbol{\beta}_c, \sigma_c^2)$ are unknown regression parameters. For a new input, this model is used to compute a probability and simulate an increasing or decreasing classification. The outcome then determines which of the following models to use to describe $r(\boldsymbol{x}, d)$ to combine with the source term.

2. For a decreasing function of distance the model is

$$logit(r(\boldsymbol{x}_i, d_i)) = \boldsymbol{h}_D(\boldsymbol{x}_i, d_i)^T \boldsymbol{\beta}_D + \varepsilon_i^D$$

3. For a non-decreasing function we assume

$$logit(K^{-1} r(\boldsymbol{x}_i, d_i)) = \boldsymbol{h}_I(\boldsymbol{x}_i, d_i)^T \boldsymbol{\beta}_I + \varepsilon_i^I.$$

Both models characterise the output relative to the assumed maximum exposure output. In model 2 the constraint is $r(\boldsymbol{x}, d) < 1$, whereas model 3 assumes $r(\boldsymbol{x}, d) < K$. The constant K is fixed for each output type. This was set as $\max_{i,d} r(\boldsymbol{x}_i, d_i)\,(1+\epsilon)$ where ϵ = 0.1 and maximisation was over all training outputs and distances. Other selections of ϵ made little difference to the results. Effectively the model says that true function ratios will be no more than 10% higher than the largest observed ratio in the data. It is also clear that $\boldsymbol{h}_I(\boldsymbol{x}, d)$ will be a non-monotonic function of distance.

The models 1-3 were fitted using the glm() function in R Version 2.14.0 to estimate parameters. Model 1 was fitted using the observed status of the 600 data functions and corresponding inputs. Models 2 and 3 were fitted using only those data functions from the respective type (increasing/decreasing). Files of coefficients and other details describing the GP model fit and GLM were stored in a file for use in separate program to generate fast predictions whenever required. Each distinct output has its own set.

2.4 Uncertainty analysis

Monte Carlo simulation was used to propagate variability in the input parameters and to quantify the impact on the output distribution. As in the Bream model, the user is able to select some of the input parameters and some of these were held fixed while others have distributions assigned to them. Details are shown in Table 1.

Input	Symbol	Possible selections or range for user input (400 training runs)	Refined range for emulation (200 training runs)	Distribution for Monte Carlo simulation (x denotes user selection)
Output_Type		Ground, IfC, IfA, AirC, AirA		
Crop_Height		0.1, 0.5, 1.0, 1.5		
N_Nozzles	x_1	12-960	24-144	
Boom_Height	x_2	0-2.5 m	0.2-1.2 m	N(mean = x, sd = 0.3x)
Forward_Speed	x_3	6-24 km/h	6-16 km/h	
Wind_Speed	x_4	1-20 m/s	1-7 m/s	N(mean = x, sd = 0.0895 + 0.1825x)
Wind_Angle	x_5	10-170°	60-120°	N(mean = x, sd =10)
Distance	d	2-20 m, user selects min and max distance		U(min, max)

Table 1. Inputs and their ranges for the approximate model

3. Results

Some examples of prediction and Monte Carlo uncertainty analysis are presented below. Here we generated 1000 simulated parameter sets, corresponding to variability in wind speed, boom height and wind angle during a spray event. The distributions are relative to nominal values that will in practice be selected by a model user (Kennedy et al, 2012). The values used here are merely for illustration. In addition, we account for variation in bystander distance by simulating a different distance uniformly between 10 and 15 m from the sprayer. The histograms show the resulting distributions of output (Figure 5).

Diagnostic plots show the 600 curves from the original model and the corresponding approximations. Typical examples are shown below for AirA1.5 (Figure 2), IfA0.1 (Figure 6), IfC1.5 (Figure 7) and Ground0.1 (Figure 8). We see here that the approximation is generally smoother than the original data, and the positive and negative slopes of the functions can be clearly seen in the predictions (notably in Figure 7 for IfC1.5). AirA is the most accurate approximation, particularly with the higher crop scenarios. The pattern of predictions broadly follows the true code outputs, and errors are low especially above 5m from the source. Errors in the predictions can be presented in various forms. Examples are shown in Figures 3, 4, 9, 10. Figure 10 gives, for selected distances from the source, fitted versus actual points for ground deposit outputs predicted at all 600 training run data. Errors are concentrated around 0, expected errors are 0 and the largest errors occur at distances closest to the source.

4. Discussion/Further work

The model described above is very fast and simple to implement. It therefore meets the requirements set out in the introduction. The R version can be readily called or adapted/recoded in other languages for use in the browse interface or different Browse work packages, thereby ensuring consistent exposure calculations under equivalent scenarios. Many thousands of MC simulations are possible in a few seconds, so that estimates of extreme exposure probabilities can be calculated in real-time.

A more detailed assessment of the accuracy of these predictions should be carried out when the scenarios are implemented. It will then be much clearer which aspects are of most importance to refine. For example, predicting highly accurate exposures at distances close to the source may be less important if no bystander is ever located there, but will probably be more important for operators. The relative errors are likely to be smaller than those shown above, when the output of interest uses multiple simulated realisations, i.e. multiple approximations of the code for different input parameters. This is because the errors will cancel out when taking an average value from those simulations. As shown, the expected error is close to zero.

More complex statistical methods are available for functional data analysis. However, these would require substantial development time as well as computation time and therefore less useful within Browse. Classifying the functions as increasing/decreasing can also be carried out in other ways, such as linear discriminant analysis. However, this requires the assumption of normality of the independent variables corresponding to a given classification. Further work could be carried out to assess improvements available with these methods.

References

Butler Ellis, M C, Miller, P C H (2010). The Silsoe Spray Drift Model: A model of spray drift for the assessment of non-target exposures to pesticides. Biosystems Engineering **107** 169-177.

Kennedy, M. C. (2005). Gaussian Emulation Machine for Sensitivity Analysis (GEM-SA). Available from http://www.tonyohagan.co.uk/academic/GEM/index.html.

Kennedy, M. C., Butler Ellis, M. C., Miller, P. C. H. (2012). BREAM: A probabilistic Bystander and Resident Exposure Assessment Model. Computers and Electronics in Agriculture 88, 63–71

Miller, PCH and Hadfield, DJ (1989). A simulation model of the spray drift from hydraulic nozzles. Journal of Agricultural Engineering Research 42, 161-170.

Morris, M. D. And Mitchell, T. J. (1995). Exploratory designs for computational experiments. J. Statistical Planning and Inference, 43, 381-402.

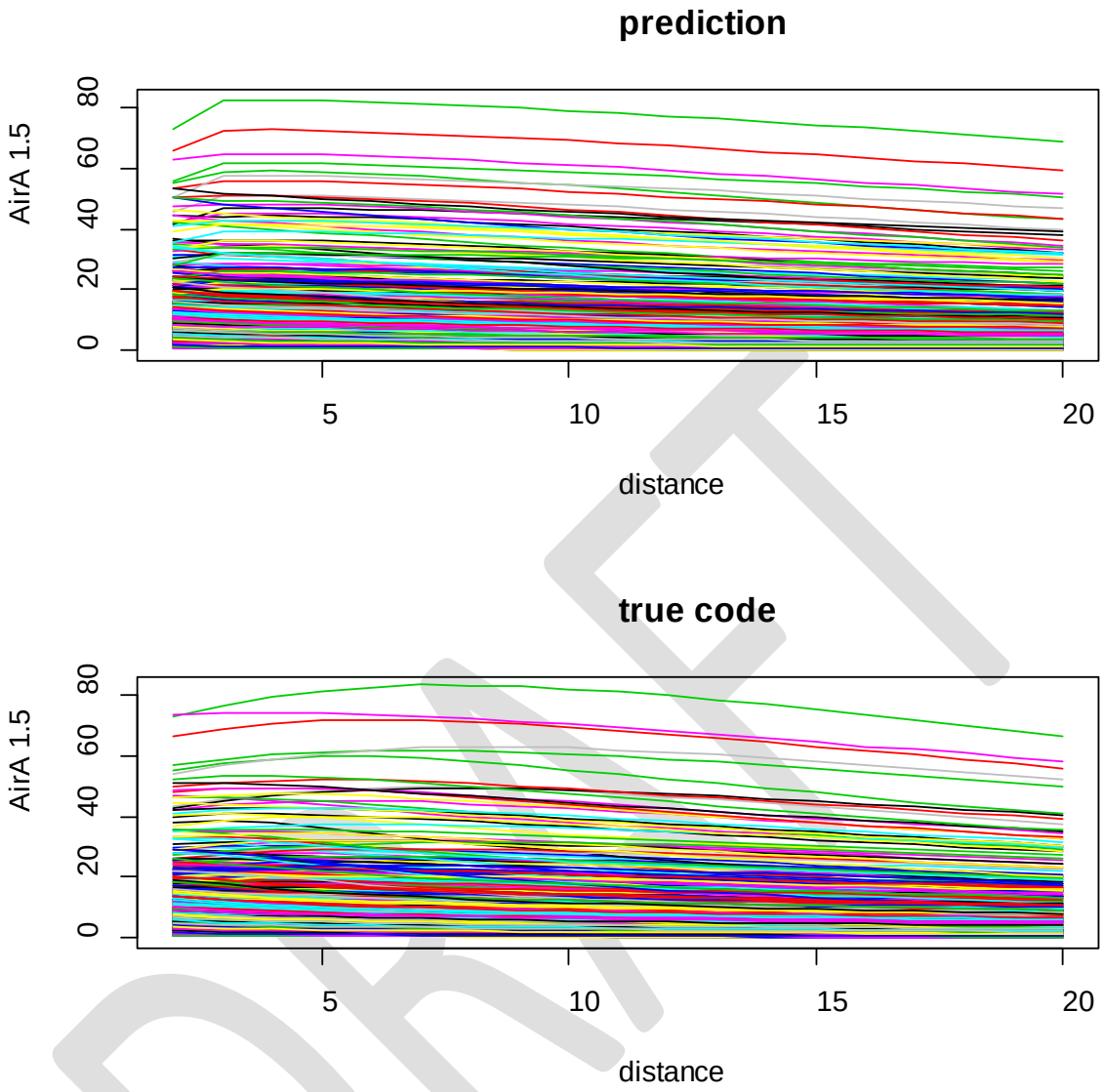

Figure 2. Comparison of Adult Airborne spray output with crop height = 1.5 m (bottom) and corresponding approximations (top) as functions of distance

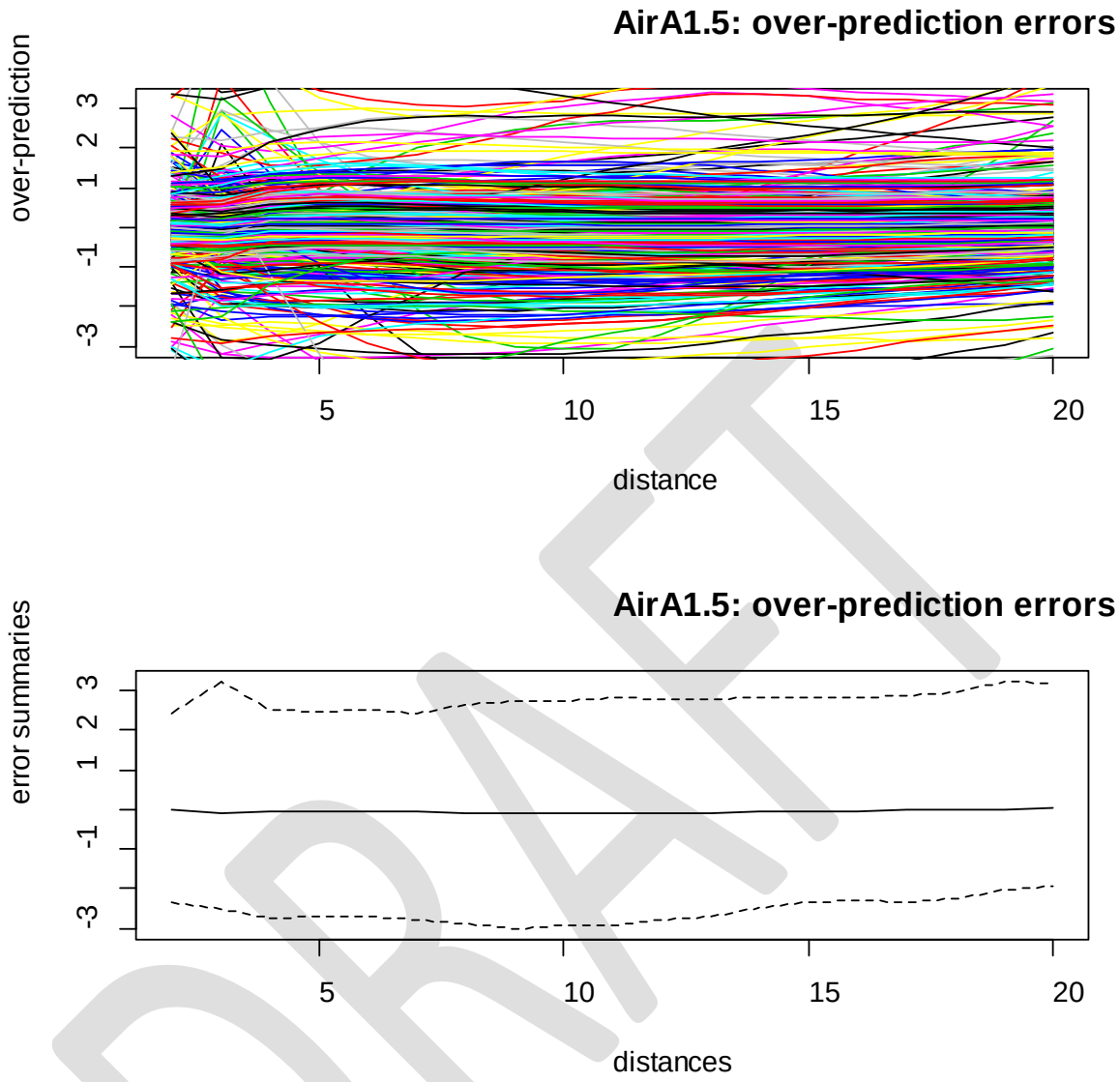

Figure 3. Cross-validation errors (as function of distance) in approximating 400 training runs for Adult Airborne spray output with crop height = 1.5 m. Solid line is median error and dashed lines are pointwise 95% credible intervals (bottom). Top panel shows errors from individual functions.

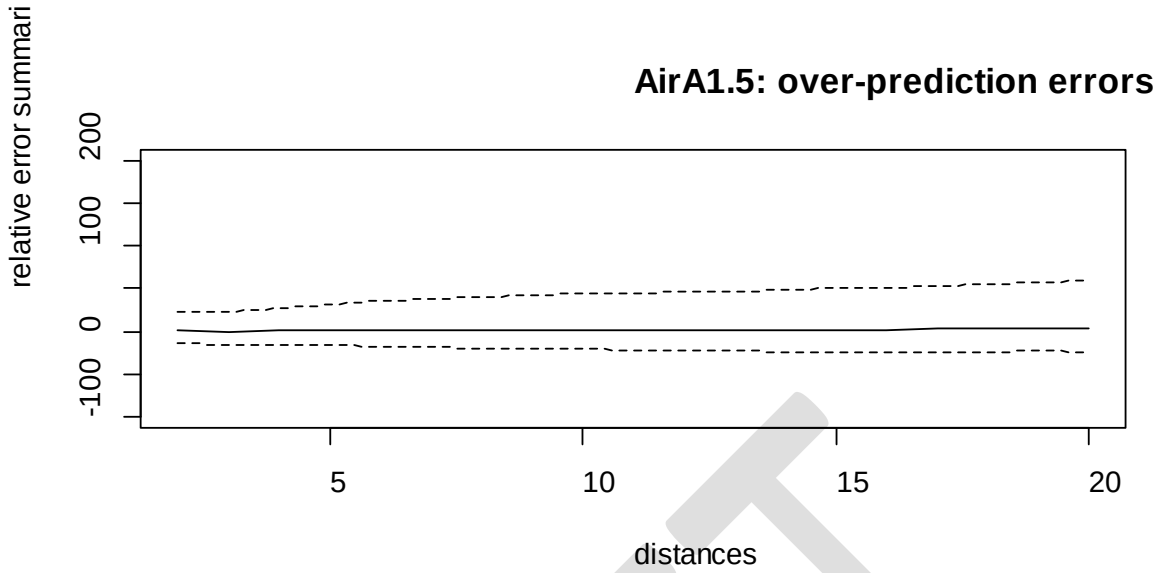

Figure 4: Relative prediction errors (as function of distance) in approximating 600 training runs for Adult Airborne spray output with crop height = 1.5 m. Solid line is median error and dashed lines are pointwise 95% credible intervals

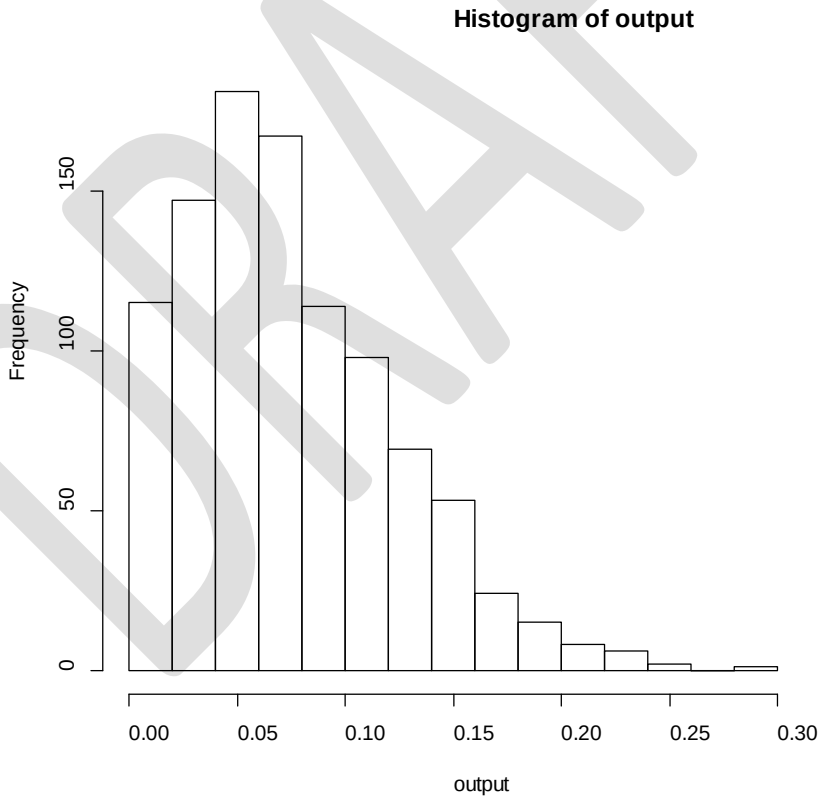

Figure 5. Distribution of Adult Inhalaton Factor for 0.1m crop height (IfA0.1) from 1000 Monte Carlo simulations and input parmeters/distributions: N_Nozzles = 100, Boom_Height ~ N(0.3, sd = 0.3 * 0.3), Forward_Speed = 4, Wind_Speed ~ N(5, 0.0068 + 0.185 * 5), Wind_Angle ~ N(10, sd = 10), distance ~ U(10, 15)

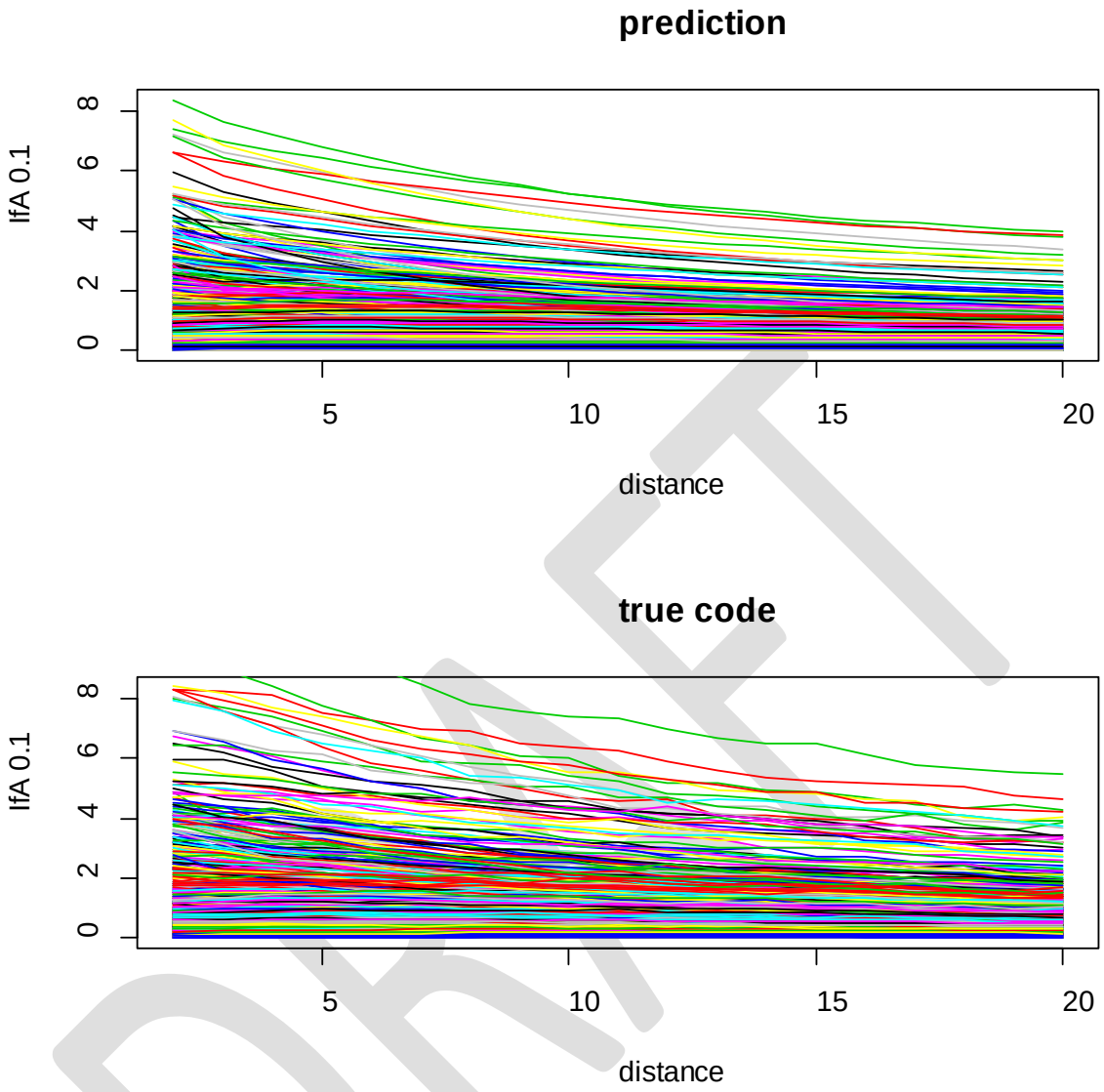

Figure 6. Predictions and true code output (as function of distance) for Adult inhalation factor with crop height = 0.1 m. All 600 code runs are plotted.

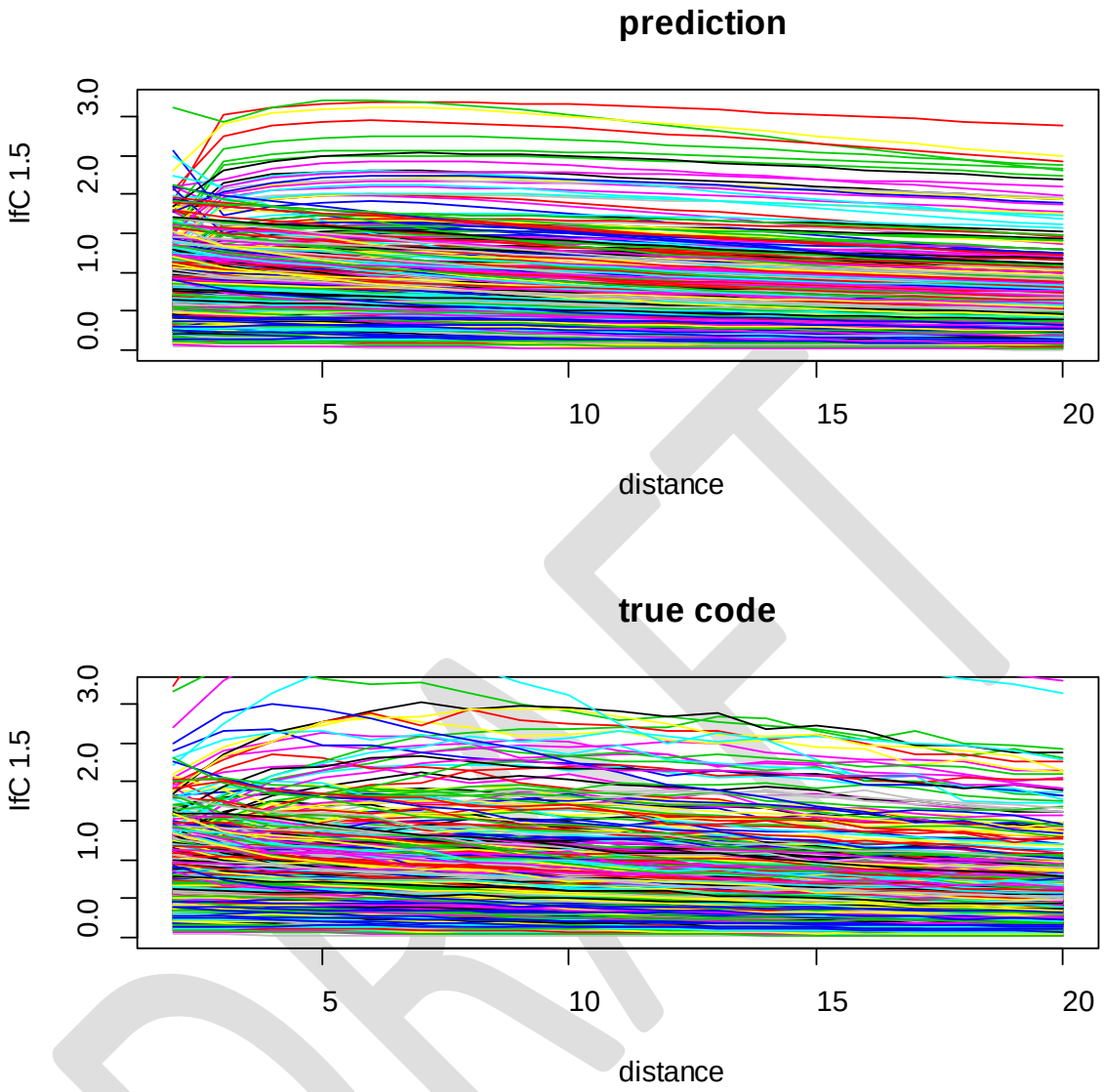

Figure 7. Predictions and true code output (as function of distance) for Child inhalation factor with crop height = 1.5 m. All 600 code runs are plotted.

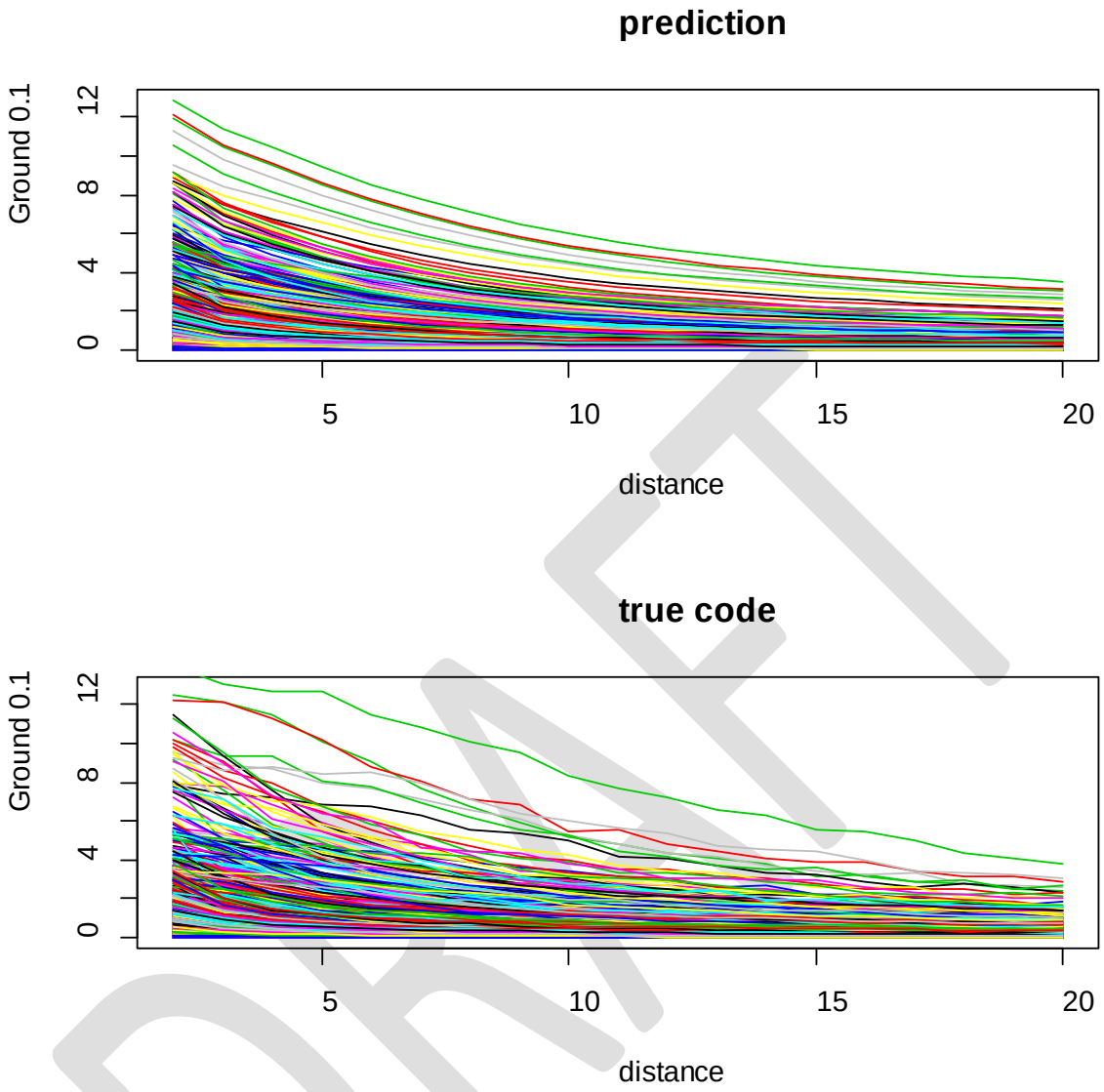

Figure 8. Predictions and true code output (as function of distance) for Ground deposits with crop height = 0.1 m. All 600 code runs are plotted.

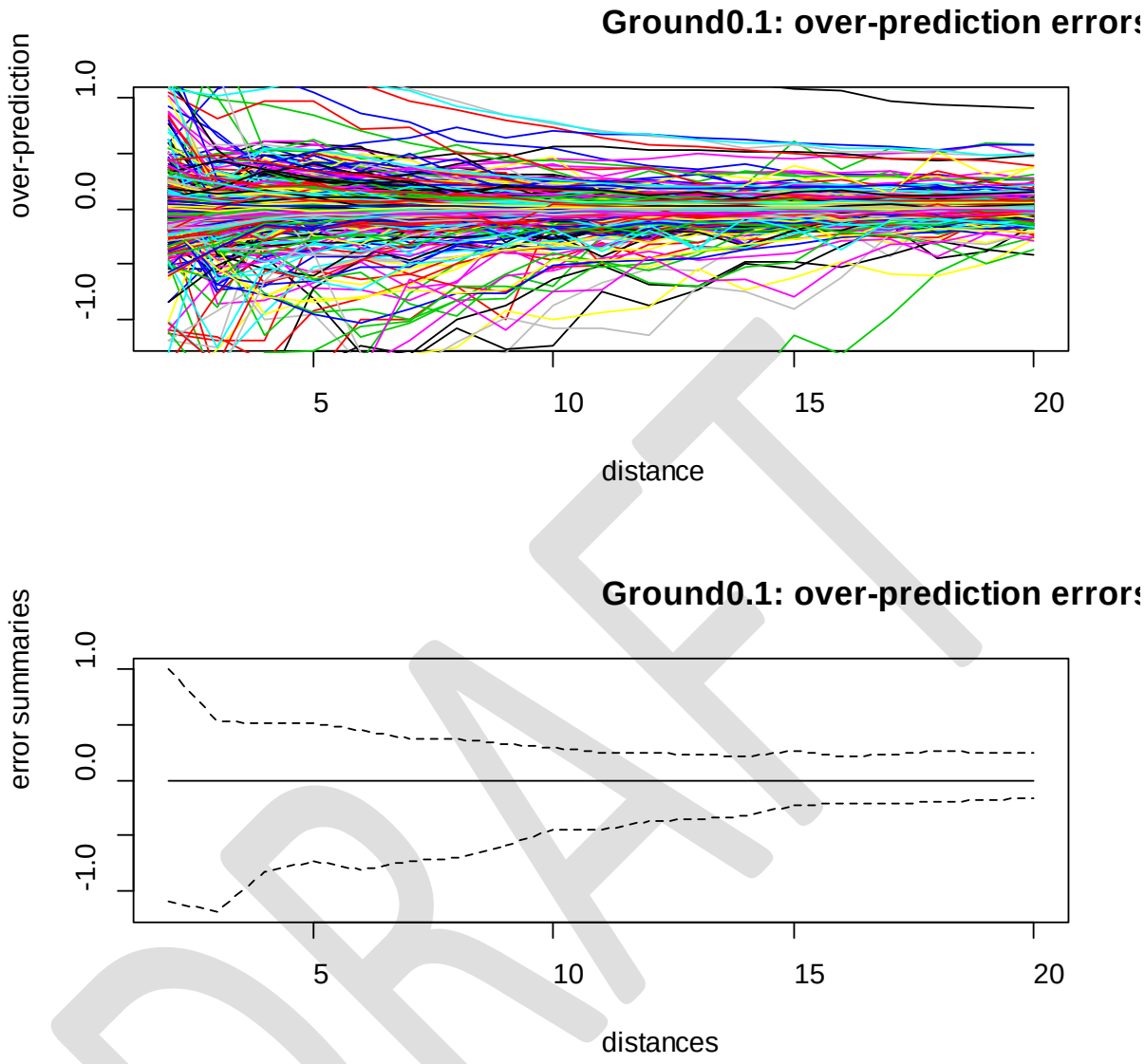

Figure 9. Prediction erros (as function of distance) for Ground deposits with crop height = 0.1 m. All 600 code runs are represented. Solid line is expected error and dotted lines are pointwise 95% credible intervals.

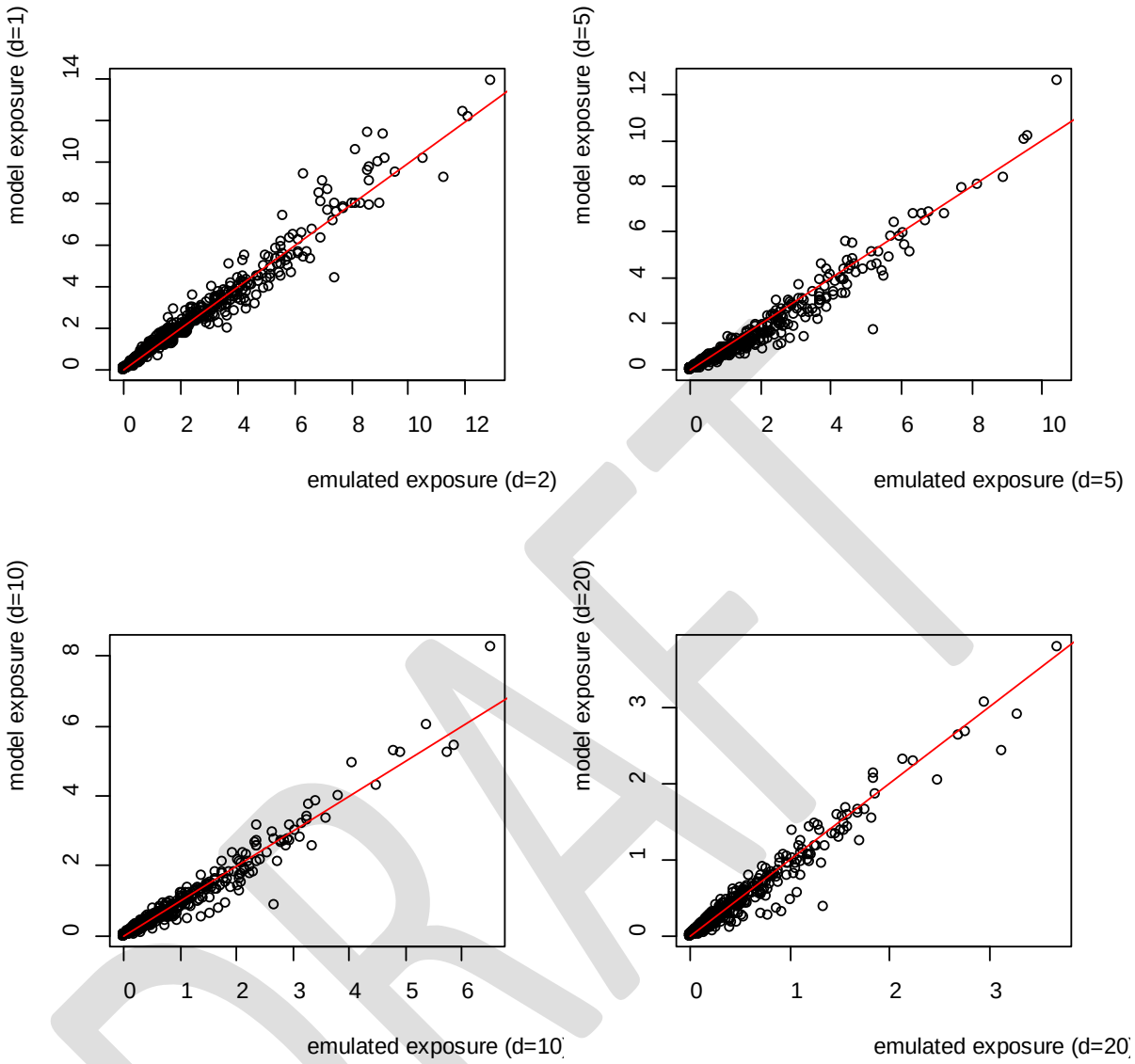

Figure 10. Fitted versus actual code outputs at distance = 2, 5, 10, 20 m for ground deposit and crop height 0.1m (Ground0.1). All 600 code runs are plotted. Red line represents perfect predictions.