

# Building kriging models using hierarchical codes with different levels of accuracy.

Loic Le Gratiet  
CEA, DAM, DIF, F-91297 Arpajon, France  
Universite Paris Diderot-Paris 7, 75205 Paris Cedex 13  
loic.le-gratiet@cea.fr

August 10, 2011

## Abstract

This paper deals with the Gaussian process based approximation of a code which can be run at different levels of accuracy using co-kriging. This method allows us to improve a surrogate model of a complex computer code using fast approximations of it. In particular, we focus on the case of large number of code levels. A thermodynamic example is used to illustrate a 3-level co-kriging.

Keywords: *co-kriging* , *multi-level code* , *computer experiment* , *surrogate models* , *Gaussian process regression*.

## 1 Introduction

Large computer codes are widely used in engineering to study physical systems since real experiments are often costly and sometimes impossible. Nevertheless, simulations can sometimes be time-consuming as well. In this case, conception based on an exhaustive exploration of the input space of the code is generally impossible under reasonable time constraints. Therefore, a mathematical approximation of the output of the code - also called surrogate model or metamodel - is often built with a few simulations to represent the real system. Gaussian Process regression is a particular class of surrogate model which makes the assumption that prior beliefs about the code can be modelled by a Gaussian Process. We focus here on this metamodel and on its extension to multiple response models. The reader is referred to [Rasmussen & Williams (2006)] for further detail about Gaussian Process models.

Actually, a computer code can often be run at different levels of complexity and a hierarchy of levels of code can hence be obtained. The aim of our research is to study the use of several levels of a code to predict the output of a costly computer code. The presented multi-stage metamodel is a particular case of co-kriging which is a well known geostatistical method.

A first metamodel for multi-level computer codes was built by [Kennedy & O'Hagan (2000)] using a spatial stationary correlation structure. Then, [Qian & Wu (2008)] built an extension to this model in a case of non spatial stationarity and [Forrester, Sobester & Keane (2007)] went into more detail about the estimation of the model parameters. Furthermore, Forrester

*et al.* presented the use of co-kriging for multi-fidelity optimization based on the EGO (Efficient Global Optimization) algorithm. A linear Bayesian approach was also proposed by [Cumming & Goldstein (2009)].

We present a new approach to estimate the parameters of the multi-level surrogate model which is effective even when many levels of code are available. Furthermore, this approach can allow us to consider prior information in the parameter estimation. We also address the problem of the co-kriging covariance matrix inversion when the number of levels is large. A solution to this problem is provided which shows that the inverse can be easily calculated. Finally, we address the problem of model validation. In particular, we present virtual cross-validation equations which give the result of the leave-one-out procedure without building sub-metamodels.

A thermodynamic example is used to illustrate a 3-level co-kriging. The purpose of this example is to predict the result of a physical experiment - which can be considered as the most costly code - modelled by an accurate computer code and by another one less accurate. The reader is referred to [Dewettinck, De Visscher, Deroo, Huyghebaert (1999)] for further detail about the example.

## 2 Example presentation: Fluidized-bed process

A fluidized-bed process is a device used in many industrial applications. In this type of process, a fluid is passed through a granular solid material at high enough velocities to suspend the solid and cause it to behave as though it were a fluid. This phenomenon is called fluidization. Fluidized-bed processes are used in the petroleum industry to produce gasoline and other fuels; they are also used in the pharmaceutical and food industries and in the water and waste treatment.

We are interested here on a particular experiment which is the measurement of the temperature of the steady-state thermodynamic operation point for a fluidized-bed process. It was presented by [Dewettinck, De Visscher, Deroo, Huyghebaert (1999)], who developed a computer model named "Topsim" to calculate the measured temperature. The code, developed for a Glatt GPCG-1 fluidized-bed unit in the top-spray configuration, can be run at 3 levels of complexity. We hence have 4 available responses:

1.  $T_{exp}$ : the experimental response.
2.  $T_3$ : the most accurate code modelling the experiment.
3.  $T_2$ : a simplified version of  $T_3$ .
4.  $T_1$ : the less accurate code modelling the experiment.

The differences between  $T_1$ ,  $T_2$  and  $T_3$  are discussed by Dewettinck et al. (1999). The aim of this study is to predict the experimental response  $T_{exp}$  given the two levels of code  $T_3$  and  $T_2$ . We only focus on a 3-level co-kriging since 28 observations are available for each level and it is not enough for a relevant 4-level co-kriging. The experimental design set and the responses  $T_1$ ,  $T_2$ ,  $T_3$  and  $T_{exp}$  are given by [Qian & Wu (2008)] who have presented a 2-level co-kriging using  $T_{exp}$  and  $T_2$ . Furthermore, the responses are parameterized by a 6-dimensional input vector presented by Dewettinck et al. (1999).

### 3 Building a model with 3 levels of code

Let assume that we have 3 levels of response,  $T_2$ ,  $T_3$  and  $T_{exp}$ . Our prior beliefs about these responses are that, given a certain set of parameters, they can be modelled by a Gaussian process. Since we have a hierarchy of 3 responses - from the less accurate to the most accurate, we can assume the autoregressive model suggested by [Kennedy & O'Hagan (2000)]:

$$T_{exp}(x) = \rho_3 T_3(x) + \delta_3(x) \quad T_3(x) \perp \delta_3(x)$$

$$T_3(x) = \rho_2 T_2(x) + \delta_2(x) \quad T_2(x) \perp \delta_2(x)$$

where:

$$\delta_3(x) \sim \mathcal{PG}(\mu_{\delta_3}, \sigma_{\delta_3}^2 r(x, x'; \theta_{\delta_3}))$$

$$\delta_2(x) \sim \mathcal{PG}(\mu_{\delta_2}, \sigma_{\delta_2}^2 r(x, x'; \theta_{\delta_2}))$$

$$T_2(x) \sim \mathcal{PG}(\mu_{T_2}, \sigma_{T_2}^2 r(x, x'; \theta_{T_2}))$$

and  $r(x, x'; \theta)$  is a correlation function with parameter  $\theta$  representing the characteristic length-scale. We note  $D_2$ ,  $D_3$  and  $D_{exp}$  the experimental design sets of  $T_2$ ,  $T_3$  and  $T_{exp}$  such that  $D_2 \subset D_3 \subset D_{exp}$ ,  $T_2^*$ ,  $T_3^*$  and  $T_{exp}^*$  are the known responses of  $T_2$ ,  $T_3$  and  $T_{exp}$  at points in  $D_2$ ,  $D_3$  and  $D_{exp}$  and  $R_{\delta_3} = r(D_{exp}, D_{exp}; \theta_{\delta_3})$ ,  $R_{\delta_2} = r(D_3, D_3; \theta_{\delta_2})$  and  $R_{T_2} = r(D_2, D_2; \theta_{T_2})$  are the correlation matrices of the different Gaussian processes. We want to determine the predictive distribution of  $T_{exp}$  given  $(T_{exp}^*, T_2^*, T_3^*, \Psi)$ , where  $\Psi = (\mu_{\delta_3}, \mu_{\delta_2}, \mu_{T_2}, \sigma_{\delta_3}, \sigma_{\delta_2}, \rho_2, \rho_3, \sigma_{T_2}, \theta_{\delta_3}, \theta_{\delta_2}, \theta_{T_2})$ . We see here that the estimation of the parameters could be an issue since they are numerous. Classical results for normal distribution give that:

$$T_{exp}(x) | T_{exp}^*, T_2^*, T_3^*, \Psi \sim \mathcal{N}(\mu(x), s^2(x))$$

where:

$$\mu(x) = \rho_2 \rho_3 \mu_{T_2} + \rho_3 \mu_{\delta_2} + \mu_{\delta_3} + t(x)^T V^{-1} M$$

$$s^2(x) = \rho_2^2 \rho_3^2 \sigma_{T_2}^2 + \rho_3^2 \sigma_{\delta_2}^2 + \sigma_{\delta_3}^2 - t(x)^T V^{-1} t(x)$$

with:

$$M = \begin{pmatrix} T_2^* - \mathbf{1}_{n_{T_2}} \mu_{T_2} \\ T_3^* - \mathbf{1}_{n_{T_3}} (\rho_2 \mu_{T_2} + \mu_{\delta_2}) \\ T_{exp}^* - \mathbf{1}_{n_{T_{exp}}} (\rho_3 \rho_2 \mu_{T_2} + \rho_3 \mu_{\delta_2} + \mu_{\delta_3}) \end{pmatrix}$$

$$t(x) = \begin{pmatrix} \rho_2 \rho_3 \sigma_{T_2}^2 r(D_2, x; \theta_{T_2}) \\ \rho_2^2 \rho_3 \sigma_{T_2}^2 r(D_3, x; \theta_{T_2}) + \rho_3 \sigma_{\delta_2}^2 r(D_3, x; \theta_{\delta_2}) \\ \rho_2^2 \rho_3^2 \sigma_{T_2}^2 r(D_{exp}, x; \theta_{T_2}) + \rho_3^2 \sigma_{\delta_2}^2 r(D_{exp}, x; \theta_{\delta_2}) + \sigma_{\delta_3}^2 r(D_{exp}, x; \theta_{\delta_3}) \end{pmatrix}$$

with  $\mathbf{1}_n$  a vector of  $n$  elements equal to 1 and  $V = \text{var}(T_2^*, T_3^*, T_{exp}^*)$ . We note that it could be an issue to invert  $V$  when the number of observations is large.

### 4 Parameter estimation and inversion of $V$

We deal in this section with the estimation of  $\Psi$  and the inversion of  $V$ . To simplify the notations, we use deterministic parameter estimation with the maximum likelihood estimate (MLE). The proposed equations can easily be used in a Bayesian approach in order to consider

prior information in the parameter estimation.

The MLE of  $(\mu_{\delta_3}, \mu_{\delta_2}, \mu_{T_2}, \sigma_{\delta_3}, \sigma_{\delta_2}, \sigma_{T_2}, \rho_2, \rho_3)$  are given by:

$$\begin{aligned}
(\hat{\mu}_{\delta_3}, \hat{\rho}_3) &= \left( \mathbf{h}_{n_{T_{exp}}}^T R_{\delta_3}^{-1} \mathbf{h}_{n_{T_{exp}}} \right)^{-1} \mathbf{h}_{n_{T_{exp}}}^T R_{\delta_3}^{-1} T_{exp}^* \\
(\hat{\mu}_{\delta_2}, \hat{\rho}_2) &= \left( \mathbf{h}_{n_{T_3}}^T R_{\delta_2}^{-1} \mathbf{h}_{n_{T_3}} \right)^{-1} \mathbf{h}_{n_{T_3}}^T R_{\delta_2}^{-1} T_3^* \\
\hat{\mu}_{T_2} &= \left( \mathbf{1}_{n_{T_2}}^T R_{T_2}^{-1} \mathbf{1}_{n_{T_2}} \right)^{-1} \mathbf{1}_{n_{T_2}}^T R_{T_2}^{-1} T_2^* \\
\hat{\sigma}_{\delta_3}^2 &= \frac{\left( T_{exp}^* - \hat{\rho}_3 T_3^*(D_{exp}) - \mathbf{1}_{n_{T_{exp}}} \hat{\mu}_{\delta_3} \right)^T R_{\delta_3}^{-1} \left( T_{exp}^* - \hat{\rho}_3 T_3^*(D_{exp}) - \mathbf{1}_{n_{T_{exp}}} \hat{\mu}_{\delta_3} \right)}{n_{T_{exp}} - 2} \\
\hat{\sigma}_{\delta_2}^2 &= \frac{\left( T_3^* - \hat{\rho}_2 T_2^*(D_3) - \mathbf{1}_{n_{T_3}} \hat{\mu}_{\delta_2} \right)^T R_{\delta_2}^{-1} \left( T_3^* - \hat{\rho}_2 T_2^*(D_3) - \mathbf{1}_{n_{T_3}} \hat{\mu}_{\delta_2} \right)}{n_{T_3} - 2} \\
\hat{\sigma}_{T_2}^2 &= \frac{\left( T_2^* - \mathbf{1}_{n_{T_2}} \hat{\mu}_{T_2} \right)^T R_{T_2}^{-1} \left( T_2^* - \mathbf{1}_{n_{T_2}} \hat{\mu}_{T_2} \right)}{n_{T_2} - 1}
\end{aligned}$$

with  $T_3^*(D_{exp})$  the responses of  $T_3^*$  at points in  $D_{exp}$ ,  $T_2^*(D_3)$  the responses of  $T_2^*$  at points in  $D_3$ ,  $\mathbf{h}_{n_{T_{exp}}} = (\mathbf{1}_{n_{T_{exp}}} \quad T_3^*(D_{exp}))$  and  $\mathbf{h}_{n_{T_3}} = (\mathbf{1}_{n_{T_3}} \quad T_2^*(D_3))$ . The closed form expression for the estimation of  $(\mu_{\delta_3}, \mu_{\delta_2}, \mu_{T_2}, \sigma_{\delta_3}, \sigma_{\delta_2}, \sigma_{T_2}, \rho_2, \rho_3)$  is an original result which is not present in the cited papers. We estimate  $(\theta_{\delta_3}, \theta_{\delta_2}, \theta_{T_2})$  by minimizing the opposite of the concentrated restricted log-likelihoods:

$$\begin{aligned}
&\log(|\det(R_{\delta_3})|) + (n_{T_{exp}} - 2)\log(\hat{\sigma}_{\delta_3}^2) \\
&\log(|\det(R_{\delta_2})|) + (n_{T_3} - 2)\log(\hat{\sigma}_{\delta_2}^2) \\
&\log(|\det(R_{T_2})|) + (n_{T_2} - 1)\log(\hat{\sigma}_{T_2}^2)
\end{aligned}$$

The parameter estimation previously presented is important since it shows that the parameter estimation for a 3-level co-kriging is equivalent to the one for 3 independent krigings. Furthermore, this result can be extended to a s-level co-kriging.

We now address the problem of the inversion of  $V$ . By sorting the experimental design sets such that  $D_3 = (D_3 \setminus D_{exp}, D_{exp})$  and  $D_2 = (D_2 \setminus D_3, D_3)$ , it can be shown that the inverse of  $V$  has the form:

$$\begin{aligned}
V^{-1} &= \begin{pmatrix} W^{-1} + \begin{pmatrix} 0 & 0 \\ 0 & \rho_3^2 \frac{R_{\delta_3}^{-1}}{\sigma_{\delta_3}^2} \end{pmatrix} & - \begin{pmatrix} 0 \\ \rho_3 \frac{R_{\delta_3}^{-1}}{\sigma_{\delta_3}^2} \end{pmatrix} \\ - \begin{pmatrix} 0 & \rho_3 \frac{R_{\delta_3}^{-1}}{\sigma_{\delta_3}^2} \end{pmatrix} & \frac{R_{\delta_3}^{-1}}{\sigma_{\delta_3}^2} \end{pmatrix} \\
W^{-1} &= \begin{pmatrix} \frac{R_{T_2}^{-1}}{\sigma_{T_2}^2} + \begin{pmatrix} 0 & 0 \\ 0 & \rho_2^2 \frac{R_{\delta_2}^{-1}}{\sigma_{\delta_2}^2} \end{pmatrix} & - \begin{pmatrix} 0 \\ \rho_2 \frac{R_{\delta_2}^{-1}}{\sigma_{\delta_2}^2} \end{pmatrix} \\ - \begin{pmatrix} 0 & \rho_2 \frac{R_{\delta_2}^{-1}}{\sigma_{\delta_2}^2} \end{pmatrix} & \frac{R_{\delta_2}^{-1}}{\sigma_{\delta_2}^2} \end{pmatrix}
\end{aligned}$$

We have here reduced the complexity of the processing of  $V^{-1}$  by deducing it from  $R_{\delta_3}^{-1}$ ,  $R_{\delta_2}^{-1}$  and  $R_{T_2}^{-1}$ . This result shows that building a 3-level co-kriging is equivalent to build 3 independent krigings. This can also be extended for a s-level co-kriging.

## 5 Model Validation

In this section, we consider how to use the cross-validation method for model validation. The idea is to split the training set into two disjoint sets, one is used for training and the other is used to monitor the performance of the model. Then, the performance on the validation set is used as a proxy for the generalization error. We present here the case of the Leave-One-Out Cross-Validation (LOO-CV) where  $n$  validation sets are obtained by removing one observation at a time. This procedure can be expensive but Rasmussen et al. [Rasmussen & Williams (2006)] show that, in the case of kriging, there are computational shortcuts. We present in this section the adaptation of these shortcuts to the case of co-kriging.

Considering  $\Psi$  as known and using the proof presented in [Rasmussen & Williams (2006)], it can be shown that the expressions for the LOO-CV predictive mean and variance at point  $x_i \in D_{exp}$  is:

$$\begin{aligned} \mu_i &= T_{exp}(x_i) - \frac{\left[ R_{\delta_3}^{-1} \left( T_{exp}^* - \hat{\rho}_3 T_3^*(D_{exp}) - \mathbf{1}_{n_{T_{exp}}} \hat{\mu}_{\delta_3} \right) \right]_i}{\left[ R_{\delta_3}^{-1} \right]_{ii}} \\ &\quad - \hat{\rho}_3 \frac{\left[ R_{\delta_2}^{-1} \left( T_3^* - \hat{\rho}_2 T_2^*(D_3) - \mathbf{1}_{n_{T_3}} \hat{\mu}_{\delta_2} \right) \right]_i}{\left[ R_{\delta_2}^{-1} \right]_{ii}} - \hat{\rho}_2 \hat{\rho}_3 \frac{\left[ R_{T_2}^{-1} \left( T_2^* - \mathbf{1}_{n_{T_2}} \hat{\mu}_{T_2} \right) \right]_i}{\left[ R_{T_2}^{-1} \right]_{ii}} \\ \sigma_i^2 &= \frac{\hat{\sigma}_{\delta_3}^2}{\left[ R_{\delta_3}^{-1} \right]_{ii}} + \hat{\rho}_3^2 \frac{\hat{\sigma}_{\delta_2}^2}{\left[ R_{\delta_2}^{-1} \right]_{ii}} + \hat{\rho}_2^2 \hat{\rho}_3^2 \frac{\hat{\sigma}_{T_2}^2}{\left[ R_{T_2}^{-1} \right]_{ii}} \end{aligned}$$

where  $A_{ii}$  is the  $(\xi_i, \xi_i)$  element of  $A$  with  $\xi_i$  representing the line number of  $x_i$  in  $A$ . Since the equations for the LOO-CV can be directly deduced from the 3-level co-kriging equations, the computational expense of computing these quantities is negligible. We note that we have also removed  $x_i$  from  $D_3$  and  $D_2$ . We can have similar equations if we decide to remove  $x_i$  only from  $D_{exp}$  or  $D_{exp}$  and  $D_3$ .

## 6 Application: Fluidized-bed process

### 6.1 Model building

The 3-level co-kriging is built with 10 measures of  $T_{exp}$ , 20 simulations of  $T_3$  and 28 simulations of  $T_2$  and we use the Matern- $\frac{5}{2}$  kernel for all the covariance functions. The estimates of  $\theta_{\delta_3}$ ,  $\theta_{\delta_2}$  and  $\theta_{T_2}$  are given in Table 1.

$\hat{\theta}_{T_2}$	0.890	0.721	2.008	2.952	1.790	0.241
$\hat{\theta}_{\delta_2}$	1.810	1.842	2.008	1.036	0.001	0.345
$\hat{\theta}_{\delta_3}$	1.790	3.988	1.218	1.790	3.595	0.722

Table 1: Estimation of the hyper-parameters (characteristic length-scale) for the 3-level co-kriging.

Furthermore, Table 2 gives the estimates of the variance and regression parameters.

$\hat{\mu}_{T_2}$	47.02
$\begin{pmatrix} \hat{\rho}_2 \\ \hat{\mu}_{\delta_2} \end{pmatrix}$	$\begin{pmatrix} 0.95 \\ 1.93 \end{pmatrix}$
$\begin{pmatrix} \hat{\rho}_3 \\ \hat{\mu}_{\delta_3} \end{pmatrix}$	$\begin{pmatrix} 0.97 \\ -0.17 \end{pmatrix}$
$\hat{\sigma}_{T_2}^2$	38.22
$\hat{\sigma}_{\delta_2}^2$	1.05
$\hat{\sigma}_{\delta_3}^2$	0.29

Table 2: Estimation of the variance and regression parameters for the 3-level co-kriging.

## 6.2 Model validation

We present in this section the results of the LOO-CV presented in Section 5. We consider two cases:

- (1) For each point  $x_i$  removed from  $D_{exp}$ , we also remove it from  $D_3$  and  $D_2$ .
- (2) For each point  $x_i$  removed from  $D_{exp}$ , we do not remove it from  $D_3$  and  $D_2$ .

Case (1) corresponds to the equations presented in Section 5 and in case (2) we have:

$$\mu_i = T_{exp}(x_i) - \frac{\left[ R_{\delta_3}^{-1} \left( T_{exp}^* - \hat{\rho}_3 T_3^*(D_{exp}) - \mathbf{1}_{n_{T_{exp}}} \hat{\mu}_{\delta_3} \right) \right]_i}{\left[ R_{\delta_3}^{-1} \right]_{ii}}$$

$$\sigma_i^2 = \frac{\hat{\sigma}_{\delta_3}^2}{\left[ R_{\delta_3}^{-1} \right]_{ii}}$$

The distinction between these two cases is important since a good performance for the LOO-CV (1) implies that the co-kriging may be effective over the entire input parameter space, while a good performance for the LOO-CV (2) only ensures that the co-kriging is effective where at least  $T_2$  is available. Figures 1 and 2 show the LOO-CV predictive errors and variances at the 10 measures of  $T_{exp}$ :

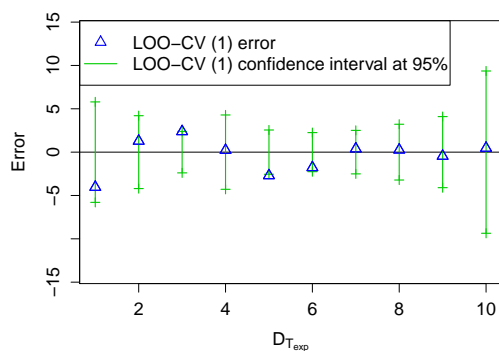


Figure 1: LOO-CV (1) prediction errors and confidence intervals at plus or minus twice the standard deviation.

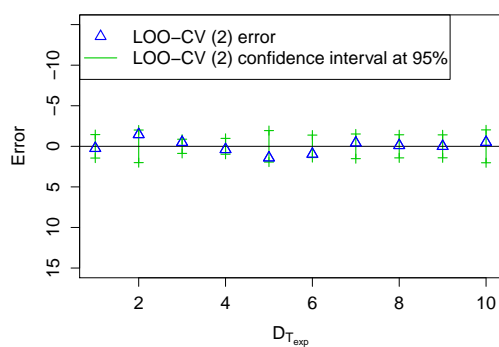


Figure 2: LOO-CV (2) prediction errors and confidence intervals at plus or minus twice the standard deviation.

Table 3 compares the two LOO-CV procedures and summarizes some results on the LOO-CV performance.

	$Q_2$	RMSE	MaxAE
LOO-CV (1)	84.01 %	1.86	4.04
LOO-CV (2)	97.32 %	0.77	1.45
	Average Std. dev.	Median Std. dev.	Maximal Std. dev
LOO-CV (1)	2.03	1.82	4.68
LOO-CV (2)	0.75	0.71	1.01

Table 3: Comparison between LOO-CV (1) and (2). Predictions are better for the LOO-CV (2) and the prediction variance seems well-evaluated since the RMSE and the average standard deviation are close.

Note that for the LOO-CV (2) the error can be important and the  $Q_2$  coefficient is not so high (it has to be close to 1). The comparison between the two LOO-CV shows that our co-kriging model is effective only where at least  $T_2$  is available. This is not surprising since only 10 measures were performed in a 6-dimensional input space. However, it highlights a strength of the proposed method, since it allows us to make good low-cost predictions requiring only runs of the unrefined code and not of the expensive one.

## 7 Conclusion

We have presented a method for building kriging models using a hierarchy of codes with different levels of accuracy. This method allows us to improve a surrogate model built on a complex code using information from cheap ones. It is particularly useful when the complex code is very expensive. An example has been provided showing the effectiveness of the method. Indeed, even when the method does not allow us to have a good surrogate model over the entire input parameter space, it can provide good predictions if a cheap version of the code is available. This could be very interesting if we want low-cost predictions.

## References

- [Rasmussen & Williams (2006)] RASMUSSEN, C. E. & WILLIAMS, C. K. I. 2006 *Gaussian Processes for Machine Learning*, the MIT Press.
- [Kennedy & O'Hagan (2000)] KENNEDY, M. C. & O'HAGAN, A. 2000 Predicting the output from a complex computer code when fast approximations are available. *Biometrika* **87**, 1-13.
- [Forrester, Sobester & Keane (2007)] FORRESTER, A. I. J., SOBESTER, A. & KEANE, A. J. 2007 Multi-fidelity optimization via surrogate modelling. *Proc. R. Soc. A* **463**, 3251-3269.
- [Qian & Wu (2008)] QIAN, Z. & JEFF WU, C. F. 2008 Bayesian Hierarchical Modeling for Integrating Low-accuracy and High-accuracy Experiments. *Technometrics* **50**, 192-204.
- [Cumming & Goldstein (2009)] CUMMING, J. A. & GOLDSTEIN, M. 2009 Small Sample Bayesian Designs for Complex High-Dimensional Models Based on Information Gained Using Fast Approximations. *Technometrics* **51**, 377-388.
- [Dewettinck, De Visscher, Deroo, Huyghebaert (1999)] DEWETTINCK, K., DE VISSCHER, A., DEROO, L. & HUYGHEBAERT, A. 1999 Modeling the steady-state thermodynamic operation point of top-spray fluidized bed processing. *Journal of Food Engineering* **39**, 131-143.