Supplementary Material to the Article: Relating Particle Properties to Biological Outcomes in Exposure Escalation Experiments

December 18, 2012

KEYWORDS: Dose-Response Models, Model Selection, Nanoinformatics, Smoothing Splines.

Appendix A: Full Conditional Distributions

In this appendix, we describe some of the full conditional distributions for the model described in the paper. Let $y_{ij}(d,t)$ denote a multivariate response corresponding to ENM i ($i = 1, ..., n$) and replicate j ($j = 1, ..., m$), at dose $d = (d_1, \ldots, d_{n_d}) \in [0, D]$ and time $t = (t_1, \ldots, t_{n_t}) \in [0, T]$. Here D is the largest measured dose and T is the largest measured exposure time. Let $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\epsilon, \boldsymbol{\sigma}_{\boldsymbol{\beta}}, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\rho})$ denote the full parameter vector, and let $\boldsymbol{\theta}_{\backslash \boldsymbol{\delta}}$ denote the vector containing all components of $\boldsymbol{\theta}$ except for some parameter δ in $\boldsymbol{\theta}$. Moreover, we denote with $\boldsymbol{y_i}$ the complete set of response values for particle i. Finally, let $\boldsymbol{h}_{dt}$ denote a $(M_d \times M_t)$-dimensional design vector, which can be defined as $(\mathcal{B}_1(d)\mathcal{B}_1(t), \ldots, \mathcal{B}_{m_d}(d)\mathcal{B}_{m_t}(t), \ldots \mathcal{B}_{M_d}(d)\mathcal{B}_{M_t}(t))$ (§2.4), and $\boldsymbol{X}$ an $n \times p$ dimensional design matrix which includes the p covariates. Using the notation above we define the full conditional distributions for all available parameters as follows.

A.1: Full conditional distributions for α_i and $\boldsymbol{\beta_i}$

Let $\eta = m \times n_d \times n_t$ be the total sample size for any particle i. Also let $y^*_{ij}(d,t) = y_{ij}(d,t) - \boldsymbol{h}'_{dt}\boldsymbol{\beta}_i$, where $\boldsymbol{\beta}_i = \mathbf{0}$ if $\gamma_i = 0$, we have

$$\alpha_i \mid \boldsymbol{y}_i, \boldsymbol{\theta}_{\backslash \alpha_i} \sim N\left(\frac{c}{c\,\eta + 1}\mathbf{1}'_\eta \boldsymbol{y}^*_{\boldsymbol{i}},\ \frac{\sigma^2_\epsilon}{\tau_i}\frac{c}{c\,\eta + 1}\right).$$

Furthermore, defining $\tilde{y}_{ij}(d,t) = y_{ij}(d,t) - \alpha_i$, we have

$$\boldsymbol{\beta_i} \mid \boldsymbol{y_i}, \gamma_i = 1, \boldsymbol{\theta}_{\backslash \boldsymbol{\beta_i}} \sim N\left(\left(\boldsymbol{\Sigma}^{\text{-}1}_{\boldsymbol{\beta_i}} + m\sum_{d,t}\frac{\boldsymbol{h}_{dt}\boldsymbol{h}'_{dt}}{\sigma^2_\epsilon/\tau_i}\right)^{\text{-}1} m\sum_{d,t}\frac{\boldsymbol{h}_{dt}\tilde{y}_{ij}(d,t)}{\sigma^2_\epsilon/\tau_i},\ \left(\boldsymbol{\Sigma}^{\text{-}1}_{\boldsymbol{\beta_i}} + m\sum_{d,t}\frac{\boldsymbol{h}_{dt}\boldsymbol{h}'_{dt}}{\sigma^2_\epsilon/\tau_i}\right)^{\text{-}1}\right),$$

where $\boldsymbol{\Sigma}_{\beta_i} = \sigma^2_{\beta_i}(K_d \otimes K_t)$.

A.2: Full conditional distributions for σ_ϵ^2 and τ_i

From A.1, let $\eta = m \times n_d \times n_t$,

$$1/\sigma_\epsilon^2 \mid \boldsymbol{y_i}, \boldsymbol{\theta}_{\setminus \sigma_\epsilon} \sim Gamma\left(a_\epsilon + \frac{n \times \eta}{2}, \frac{1}{2}\sum_{d,t,j,i}\Big(y_{ij}(d,t) - m_i(d,t)\Big)^2\tau_i + b_\epsilon\right),$$

where

$$m_i(d,t) = \begin{cases} \alpha_i & \text{if } \gamma_i = 0 \\ \boldsymbol{h}'_{dtl}\boldsymbol{\beta}_i + \alpha_i & \text{if } \gamma_i = 1. \end{cases}$$

For each particle i, $(i = 1, \ldots, n)$, the variance inflation parameter τ_i is

$$\tau_i \mid \boldsymbol{y_i}, \boldsymbol{\theta}_{\setminus \tau_i} \sim Gamma\left(\frac{\nu + \eta}{2}, \sum_{d,t,j}\frac{\Big(y_{ij}(d,t) - m_i(d,t)\Big)^2}{2\sigma_\epsilon^2} + \frac{\nu}{2}\right),$$

where $m_i(d,t)$ is defined as before.

A.3: Full conditional distributions for other variance parameters

$$1/\sigma_{\beta_i}^2 \mid \boldsymbol{\theta}_{\setminus \sigma_{\beta_i}} \sim Gamma\left(a_{\beta_i} + \frac{M_d M_t}{2},\ b_{\beta_i} + \frac{1}{2}\boldsymbol{\beta}'_i(K_d \otimes K_t)\boldsymbol{\beta}_{ij}\right).$$

A.4: Full conditional distributions for $\boldsymbol{\lambda}_p$ and z_i

The latent probit scores have conditional distribution:

$$z_i \mid \boldsymbol{\lambda}, \boldsymbol{\rho}, \gamma_i = 1 \sim N\left(\boldsymbol{x}'_{i\rho}\boldsymbol{\lambda}_\rho,\ 1\right) I(z_i \leq 0), \quad z_i \mid \boldsymbol{\lambda}, \boldsymbol{\rho}, \gamma_i = 0 \sim N\left(\boldsymbol{x}'_{i\rho}\boldsymbol{\lambda}_\rho,\ 1\right) I(z_i > 0).$$

Similarly, regression coefficients $\boldsymbol{\lambda}_\rho$ are

$$\boldsymbol{\lambda}_\rho \mid \boldsymbol{z}, \boldsymbol{\rho} \sim N\left(\frac{g_\rho}{g_\rho + 1}(\boldsymbol{X}'_\rho \boldsymbol{X}_\rho)^{\text{-}1} X'_\rho \boldsymbol{z},\ \frac{g_\rho}{g_\rho + 1}(\boldsymbol{X}'_\rho \boldsymbol{X}_\rho)^{\text{-}1}\right).$$

0.1 A.5: Derivation of $p(\boldsymbol{\gamma} \mid \boldsymbol{y}, \boldsymbol{\theta}_{\setminus\boldsymbol{\gamma}})$

Let $\tilde{y}_{ij}(d,t) = y_{ij}(d,t) - \alpha_i$ and $\boldsymbol{H}_i$ be a $\eta \times M_d M_t$ matrix of tensor product spline bases. Finally, define $\boldsymbol{\Omega}_{\beta_i} = \left(\frac{\tau_i}{\sigma_\epsilon^2}\boldsymbol{H}_i'\boldsymbol{H}_i + \boldsymbol{\Sigma}_\beta^{-1}\right)$ For each particle i (i=1,...,n), we have

$$p(\gamma_i \mid \tilde{\boldsymbol{y}}_{\boldsymbol{i}}, \boldsymbol{\theta}_{\setminus\boldsymbol{\gamma}}) \;\propto\; p\Big(\tilde{\boldsymbol{y}}_{\boldsymbol{i}} \mid \boldsymbol{\theta}_{\setminus\boldsymbol{\beta_i}}\Big) p(\gamma_i \mid \boldsymbol{\lambda}, \boldsymbol{\rho}),$$

where the likelihood, marginalized with respect to $\boldsymbol{\beta}_i$, is

$$p\Big(\tilde{\boldsymbol{y}}_{\boldsymbol{i}} \mid \gamma_i = 0, \boldsymbol{\theta}_{\setminus\boldsymbol{\beta_i}\cup\boldsymbol{\gamma}}\Big) \propto \exp\left\{-\frac{\tau_i}{2\sigma_\epsilon^2}\tilde{\boldsymbol{y}}_{\boldsymbol{i}}'\tilde{\boldsymbol{y}}_{\boldsymbol{i}}\right\},$$

and

$$\begin{aligned}
p\Big(\tilde{\boldsymbol{y}}_{\boldsymbol{i}} \mid \gamma_i = 1, \boldsymbol{\theta}_{\setminus\boldsymbol{\beta_i}\cup\boldsymbol{\gamma}}\Big) &= \int p(\tilde{\boldsymbol{y}}_{\boldsymbol{i}} \mid \boldsymbol{\beta}_i, \tau_i, \sigma_\epsilon^2, \gamma_i = 1)\, p(\boldsymbol{\beta}_i \mid \boldsymbol{\Sigma}_\beta, \gamma_i = 1)\, d\boldsymbol{\beta}_i, \\
\propto\;\; & |\boldsymbol{\Sigma}_\beta|^{-\frac{1}{2}} \exp\left\{-\frac{\tau_i}{2\sigma_\epsilon^2}\tilde{\boldsymbol{y}}_{\boldsymbol{i}}'\tilde{\boldsymbol{y}}_{\boldsymbol{i}}\right\} \int \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_i'\Big(\boldsymbol{H}_i'\boldsymbol{H}_i\frac{\tau_i}{\sigma_\epsilon^2} + \boldsymbol{\Sigma}_\beta^{-1}\Big)\boldsymbol{\beta}_i - 2\frac{\tau_i}{\sigma_\epsilon^2}\boldsymbol{\beta}_i'\boldsymbol{H}_i'\tilde{\boldsymbol{y}}_{\boldsymbol{i}}\right)\right\} d\boldsymbol{\beta}_i \\
\propto\;\; & |\boldsymbol{\Sigma}_\beta|^{-\frac{1}{2}} |\boldsymbol{\Omega}_{\beta_i}|^{-\frac{1}{2}} \exp\left\{\frac{\tau_i^2}{2\sigma_\epsilon^4}\tilde{\boldsymbol{y}}_{\boldsymbol{i}}'\boldsymbol{H}_i\boldsymbol{\Omega}_{\beta_i}^{-1}\boldsymbol{H}_i'\tilde{\boldsymbol{y}}_{\boldsymbol{i}}\right\} \exp\left\{-\frac{\tau_i}{2\sigma_\epsilon^2}\tilde{\boldsymbol{y}}_{\boldsymbol{i}}'\tilde{\boldsymbol{y}}_{\boldsymbol{i}}\right\};
\end{aligned}$$

which gives the result in the manuscript (§3.2).

Appendix B: Model Assessment

In this appendix we discuss model assessment. First we assess goodness of fit using the conditional predictive ordinate (cpo), as described by Geisser (1980). Next we plot the probability integral transform (PIT) histogram, as a measure of predictive performance (Gneiting et al. (2007)). Finally, we present some graphical posterior predictive checks.

B.1: Conditional Predictive Ordinate (CPO)

The conditional predictive ordinate (CPO) is a diagnostic tool for detecting observations with poor model fit. If we let $\boldsymbol{Y}$ denote the complete set of responses, let $\boldsymbol{Y}_{-k}$ denote

observation $\boldsymbol{Y}$ with the k-th component omitted, and let $\boldsymbol{Y}_k^{obs}$ denote the kth component of observation $\boldsymbol{Y}$, then CPO_k can be defined as follows:

$$\begin{aligned} CPO_k &= \pi(\boldsymbol{Y}_k^{obs} \mid \boldsymbol{Y}_k) = \int \pi(\boldsymbol{Y}_k^{obs} \mid \boldsymbol{Y}_{-k}, \boldsymbol{\omega}) \pi(\boldsymbol{\omega} \mid \boldsymbol{Y}_{-k}) \, \mathrm{d}\boldsymbol{\omega}, \\ \pi(\boldsymbol{Y}_k^{obs} \mid \boldsymbol{Y}_{-k}, \boldsymbol{\omega}) &= \frac{1}{\int \frac{\pi(\boldsymbol{\omega} \mid \boldsymbol{Y})}{\pi(\boldsymbol{Y}_{\boldsymbol{k}}^{\boldsymbol{obs}} \mid \boldsymbol{\omega})} \, \mathrm{d}\boldsymbol{\omega}}. \end{aligned} \tag{1}$$

Here, $\boldsymbol{\omega} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}, \sigma_\epsilon, \boldsymbol{\sigma}_\beta, \boldsymbol{\lambda}, \boldsymbol{\gamma}, \boldsymbol{\rho})$ denotes the full parameter vector. Given N MCMC samples, $n = 1, ..., N$, from the posterior distribution $P(\boldsymbol{\omega} \mid \boldsymbol{Y})$, we can obtain the harmonic mean estimate of CPO_k as follows:

$$\hat{CPO}_k = \frac{N}{\sum_{n=1}^{N} 1/\pi\left(\boldsymbol{Y}_{\boldsymbol{k}}^{\boldsymbol{obs}} | \boldsymbol{\omega}_k^{(n)}\right)}. \tag{2}$$

The expression above is evaluated at posterior samples $\boldsymbol{\omega}_k^{(1)}, \ldots, \boldsymbol{\omega}_k^{(N)}$.

A plot of -log(CPO_k) can be used to diagnose poor model fit. Large values of -log(CPO_k) indicate observations that are not consistent with the model. The *top* panel of Figure 1, provides a plot of -log($\hat{CPO}_i(d,t)$) for the model and data described in the main article. Overalll values of -log($\hat{CPO}_i(d,t)$) are relatively low, indicating good model fit. The *middle* panel indicates that the largest values of -log($\hat{CPO}_i(d,t)$) tend to be observations with large exposure times, This is to be expected, as cell death is followed after sometime by the dissolution of cell nuclei, hindering the measurement of cellular responses.

B.2: Probability Integral Transform (PIT)

The probability integral transform (PIT), as described by Gneiting et al. (2007), is frequently used as a measure of posterior predictive calibration. Here calibration is defined as the statistical consistency between the posterior predictive distribution and the observed responses $\boldsymbol{Y}$. The PIT is described as the value of the observed response $\boldsymbol{Y_k}$ attained under the predictive cumulative distribution function. Using the same notation as above, the PIT can be

defined as follows:

$$PIT_k = \int P(\boldsymbol{Y}_k \leq \boldsymbol{\omega})\pi(\boldsymbol{\omega})\mathrm{d}\boldsymbol{\omega} = P_k(\boldsymbol{Y}_k). \tag{3}$$

Given N MCMC samples, $n = 1, ..., N$, from the posterior distribution $P(\boldsymbol{\omega} \mid \boldsymbol{Y})$, we can estimated PIT as follows:

$$\hat{PIT}_k = \tfrac{1}{N} \sum_{n=1}^{N} I(\boldsymbol{Y_k} \leq \tilde{\boldsymbol{Y}}_k^{(n)}). \tag{4}$$

where $\boldsymbol{Y}_{rep\,k}^{(n)}$ is a sample from the posterior predictive distribution.

A plot of the PIT histogram can be used to visually assess the calibration of the model. Under good predictive performance of the model, the PIT histogram has a uniform distribution (see Diebold et al. (1997) for a formal proof). Inspection of the PIT histogram can also indicate reasons for poor predictive performance. A hump-shaped PIT histogram indicates prediction intervals that are, on average, to wide due to over dispersion of the predicative distribution. A U-shaped PIT histogram indicates that the predictive distribution is too narrow. Finally, a triangle shaped PIT histogram corresponds to biased predictive distributions (Gneiting et al. 2007).

The *bottom* panel of Figure 1 provides a plot of the PIT histogram for the entire model, including all doses, times, and particles. Visual assessment indicates that the plot does tend toward uniformity, indicating good overall predictive performance.

B.3: Posterior Predictive Diagnostics

A common tool for model checking in Bayesian inference involves posterior predictive checks. The basic idea behind posterior predictive checking is that if the model is a good fit to the data, then data replicated under the model should resemble the observed response $\boldsymbol{Y}$. In posterior predictive checking, replicate samples $\boldsymbol{Y_{rep}}$, are simulated from the posterior predictive distribution and compared to the observed data $\boldsymbol{Y}$. Potential problems with the model can be detected by looking for systematic differences between the simulated posterior

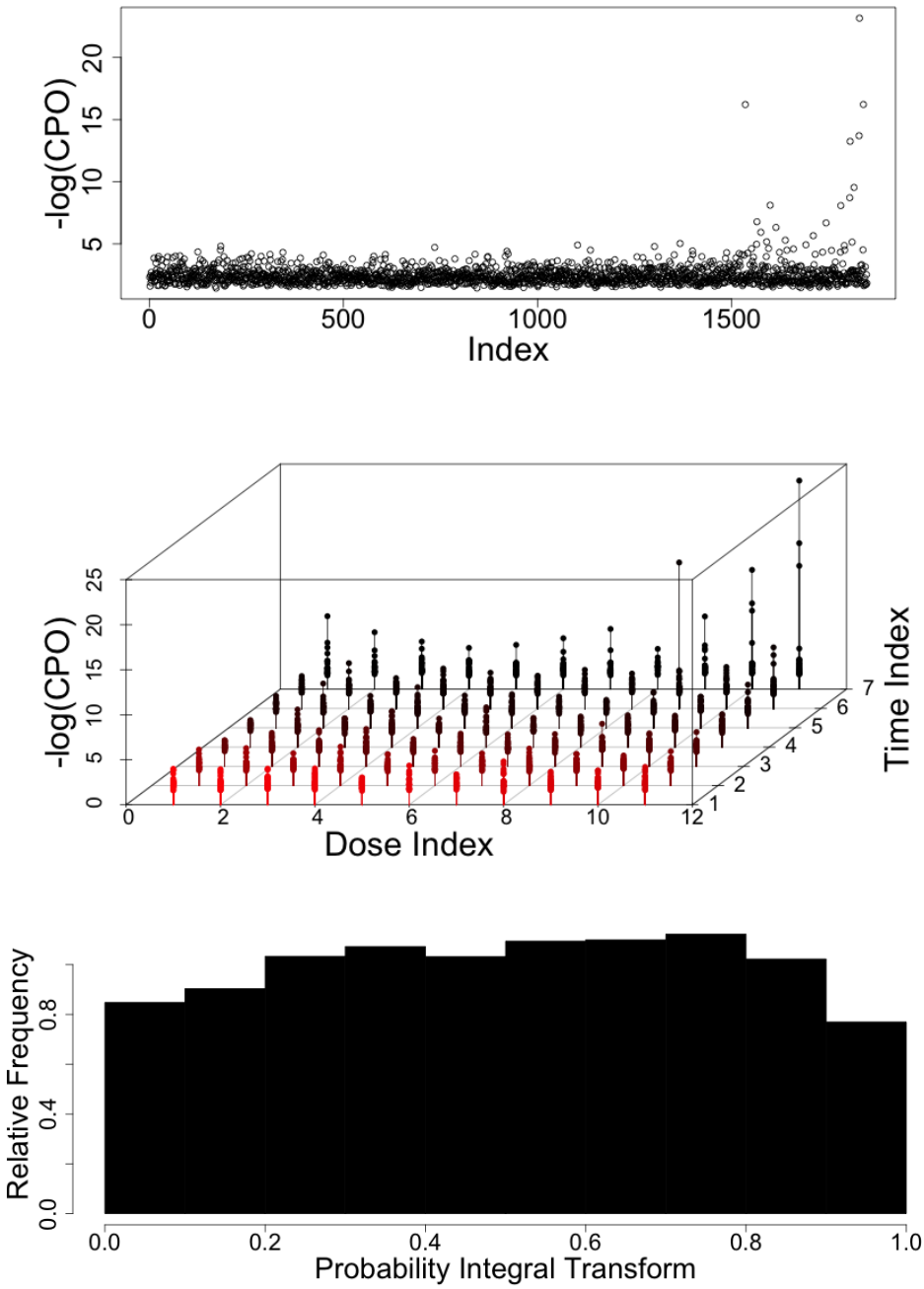

Figure 1: **Graphical model diagnostics.** *(Top)* Estimate of $-\log(cpo_i(d,t))$ for detecting observations with poor model fit. *(Middle)* Plot of $-log(cpo_i(d,t))$ as a function of dose and time, indicating any relationship between outlying observations and the administered dose or duration of exposure. *(Bottom)* Probability Integral Transform assessing empirical calibration of the posterior predictive distribution.

predictive samples and the observed response. Using the same notation described above, the posterior predictive distribution can be described as follows:

$$p(\boldsymbol{Y}_{rep} \mid \boldsymbol{Y}) = \int P(\boldsymbol{Y}_{rep} \mid \boldsymbol{\omega}) P(\boldsymbol{\omega} \mid \boldsymbol{Y}) \mathrm{d}\boldsymbol{\omega}. \tag{5}$$

Given N MCMC samples, $n = 1, ..., N$, from the posterior distribution $P(\boldsymbol{\omega} \mid \boldsymbol{Y})$, we can draw samples $\boldsymbol{Y}_{rep}^{(n)}$, $n = 1, ..., N$, from the posterior predictive distribution

Diagnostics of posterior predictive performance are obtained by comparing draws from the posterior predictive distribution to the observed data, using both formal tests and graphical checks. Graphical model checking involves the display of the simulated data from the posterior predictive distribution alongside the observed data $\boldsymbol{Y}$, and visually looking for large discrepancies such as lack of coverage (Gelman et al. 2004).

Figures 3 and 4 provide plots of the distribution of the posterior predictive mean response averaged across all doses and times of exposure (black), for each particle. The mean and associated 95% posterior intervals for the posterior predictive distribution are marked using vertical lines (black). Also included is the empirical mean response across all doses and times of exposure (red). Figure 2, summarizes these results by plotting the mean and 95% posterior intervals of the posterior predictive mean response (black), along with the the empirical mean response across all doses and times (red), for each particle. In all cases the empirical mean response is contained within the 95% posterior intervals of the posterior predictive mean distribution, indicating relatively good posterior coverage across all doses and times of exposure.

References

Diebold, F., T. Gunther, and A. Tay (1997). Evaluating density forecasts. *International Economic Review 39*, 863–883.

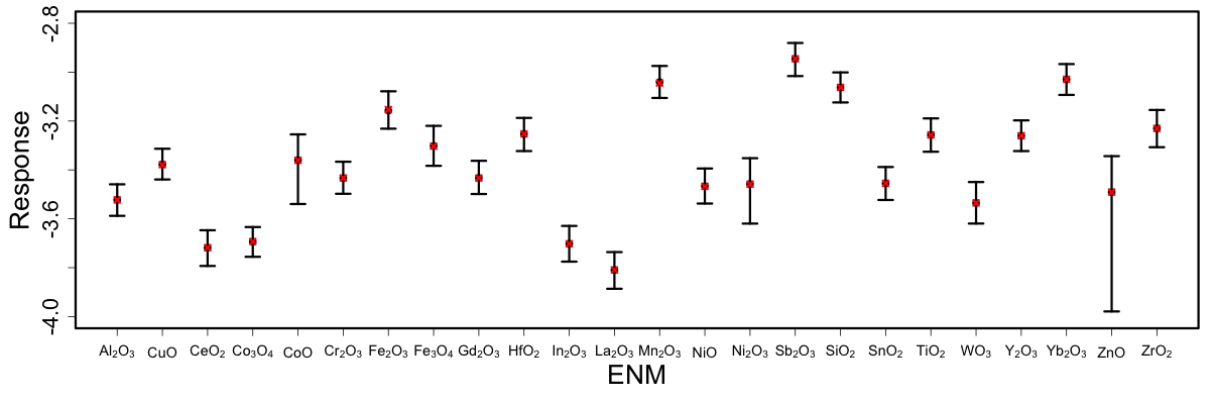

Figure 2: **Summary of posterior predictive mean coverage.** Mean and 95% posterior intervals of the posterior predictive mean response across all doses and times of exposure, for all 24 particles.) Also included are the empirical mean responses across all doses and times of exposure (red).

Geisser, S. (1980). Discussion on sampling and bayes inference in scientific modeling and robustness. *Journal of the Royal Statistical Society: Series A 143*, 416–417.

Gelman, A., J. Carlin, H. Stern, and D. Rubin (2004). *Bayesian Data Analysis*. Boca Raton, London, New York, Washington, D.C.: Chapman and Hall/CRC.

Gneiting, T., F. Balabdaoui, and A. Raftery (2007). Probabilistic forecasts, calibration and sharpness. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(2), 243–268.

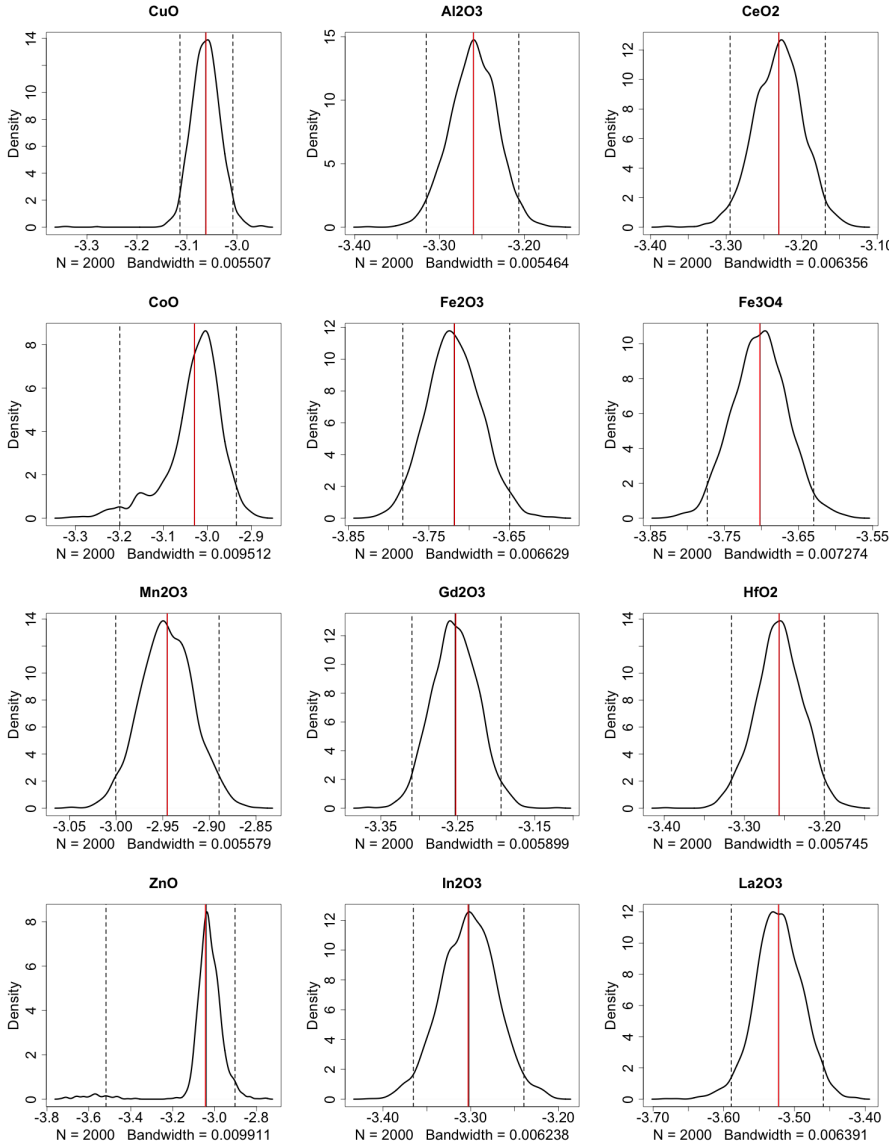

Figure 3: **Posterior predictive mean distributions for CuO, Al_2O_3, CeO_2, CoO, Fe_2O_3, Fe_3O_4, Mn_2O_3, Gd_2O_3, HfO_2, ZnO, In_2O_3, and La_2O_3 ENMs.** For each particle we plot the distribution of the posterior predictive mean response across all doses and times of exposure (black), along with the mean (solid black line) and associated 95% posterior intervals (dotted black lines) for this distribution. Also included is the empirical mean response across all doses and times of exposure (red).

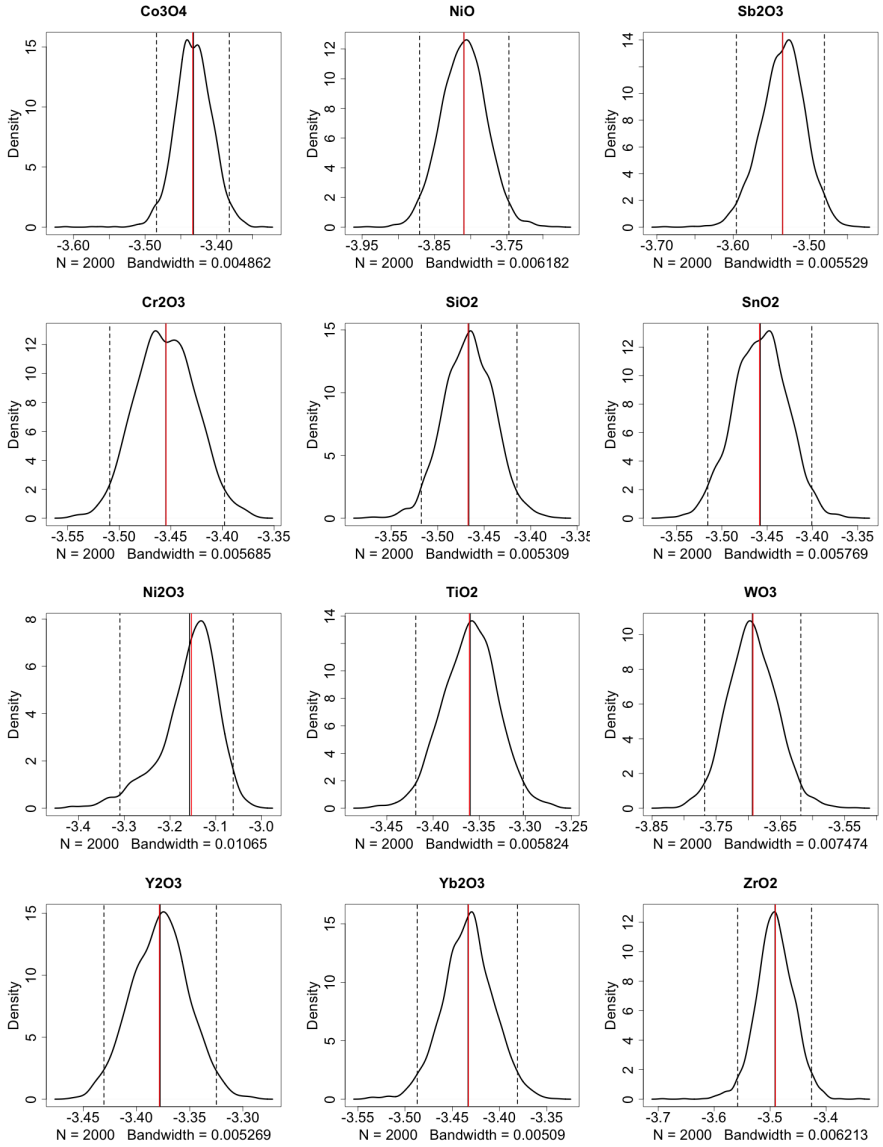

Figure 4: **Posterior predictive mean distributions for Co_3O_4, NiO, Sb_2O_3, Cr_2O_3, SiO_2, SnO_2, Ni_2O_3, TiO_2, WO_3, Y_2O_3, Yb_2O_3, and ZrO_2 ENMs.**] For each particle we plot the distribution of the posterior predictive mean response across all doses and times of exposure (black), along with the mean (solid black line) and associated 95% posterior intervals (dotted black lines) for this distribution. Also included is the empirical mean response across all doses and times of exposure (red).