ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2019)
182, *Part* 2, *pp.* 389–402

Visualization in Bayesian workflow

Jonah Gabry,

Columbia University, New York, USA

Daniel Simpson,

University of Toronto, Canada

Aki Vehtari,

Aalto University, Espoo, Finland

Michael Betancourt

Columbia University, New York, and Symplectomorphic, New York, USA

and Andrew Gelman

Columbia University, New York, USA

[*Read before* The Royal Statistical Society *at a meeting on 'Data visualization' at the Society's 2018 annual conference in Cardiff on Wednesday, September 5th, 2018, the President*, Professor D. J. Spiegelhalter, *in the Chair*]

Summary. Bayesian data analysis is about more than just computing a posterior distribution, and Bayesian visualization is about more than trace plots of Markov chains. Practical Bayesian data analysis, like all data analysis, is an iterative process of model building, inference, model checking and evaluation, and model expansion. Visualization is helpful in each of these stages of the Bayesian workflow and it is indispensable when drawing inferences from the types of modern, high dimensional models that are used by applied researchers.

Keywords: Bayesian data analysis; Statistical graphics; Statistical workflow

1. Introduction and running example

Visualization is a vital tool for data analysis, and its role is well established in both the exploratory and the final presentation stages of a statistical workflow. In this paper, we argue that the same visualization tools should be used at all points during an analysis. We illustrate this thesis by following a single real example, estimating the global concentration of a certain type of air pollution, through all of the phases of statistical workflow:

(a) exploratory data analysis to aid in setting up an initial model;
(b) computational model checks using fake data simulation and the prior predictive distribution;
(c) computational checks to ensure that the inference algorithm works reliably;
(d) posterior predictive checks and other juxtapositions of data and predictions under the fitted model;

Address for correspondence: Jonah Gabry, Columbia University, 927 Social Work Building, 1255 Amsterdam Avenue, New York, NY 10027, USA.
E-mail: jonah.sol.gabry@columbia.edu

 0964–1998/19/182389

(e) model comparison via tools such as cross-validation.

The tools that are developed in this paper are implemented in the `bayesplot` R package (Gabry, 2017; R Core Team, 2017), which uses `ggplot2` (Wickham, 2009) and is linked to—though not dependent on—Stan (Stan Development Team, 2017a, b): the general purpose Hamiltonian Monte Carlo (HMC) engine for Bayesian model fitting.

To discuss better the ways that visualization can aid a statistical workflow we consider a particular problem: the estimation of human exposure to air pollution from particulate matter measuring less than 2.5 μm in diameter, $PM_{2.5}$. Exposure to $PM_{2.5}$ is linked to a number of poor health outcomes, and a recent report estimated that $PM_{2.5}$ is responsible for 3 million deaths world wide each year (Shaddick *et al.*, 2018).

For our running example, we use the data from Shaddick *et al.* (2018), aggregated to the city level, to estimate concentrations of ambient $PM_{2.5}$ across the world. The statistical problem is that we have direct measurements of $PM_{2.5}$ from only a sparse network of 2980 ground monitors with heterogeneous spatial coverage (Fig. 1(a)). This monitoring network has especially poor coverage across Africa, central Asia and Russia.

To estimate the public health effect of $PM_{2.5}$, we need estimates of its concentration at the same spatial resolution as the population data. To obtain these estimates, we supplement the direct measurements with a high resolution satellite data product that converts measurements of aerosol optical depth into estimates of $PM_{2.5}$ concentration. The hope is that we can use the ground monitor data to calibrate the approximate satellite measurements and hence obtain estimates of $PM_{2.5}$ concentration at the required spatial resolution.

The aim of this analysis is to build a predictive model of $PM_{2.5}$ with appropriately calibrated prediction intervals. We shall not attempt a full analysis of these data, which was undertaken by Shaddick *et al.* (2018). Instead, we shall focus on three simple, but plausible, models for the data to show how visualization can be used to help to construct, sense-check, compute and evaluate these models.

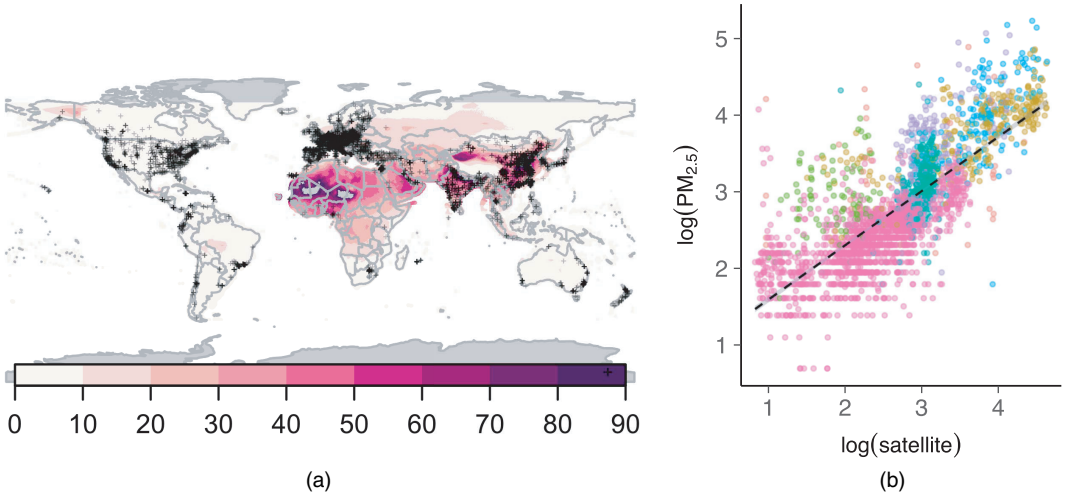

Fig. 1. Data displays for our running example of exposure to particulate matter: (a) satellite estimates of $PM_{2.5}$ concentration (●, locations of the ground monitors); (b) scatter plot of log($PM_{2.5}$) *versus* log(satellite) (●, eastern Europe–central Europe–central Asia; ●, high income super-region; ●, Latin America–Caribbean; ●, north Africa–Middle East; ●, south Asia; ●, south-east Asia–east Asia–Oceania; ●, sub-Saharan Africa)

The data that are analysed in the paper and the programs that were used analyse them can be obtained from

```
https://rss.onlinelibrary.wiley.com/hub/journal/1467985x/series-a-datasets
```

2. Exploratory data analysis goes beyond just plotting the data

An important aspect of formalizing the role of visualization in exploratory data analysis is to place it within the context of a particular statistical workflow. In particular, we argue that exploratory data analysis is more than simply plotting the data. Instead, we consider it a method to build a network of increasingly complex models that can capture the features and heterogeneities in the data (Gelman, 2004).

This ground-up modelling strategy is particularly useful when the data that have been gathered are sparse or unbalanced, as the resulting network of models is built knowing the limitations of the design. A different strategy, which is common in machine learning, is to build a top-down model that throws all available information into a complicated non-parametric procedure. This works well for data that are a good representation of the population of interest but can be prone to overfitting or generalization error when used on sparse or unbalanced data. Using a purely predictive model to calibrate the satellite measurements would yield a fit that would be dominated by data in western Europe and north America, which have air pollution profiles that are very different from those of most developing nations. With this in mind, we use the ground-up

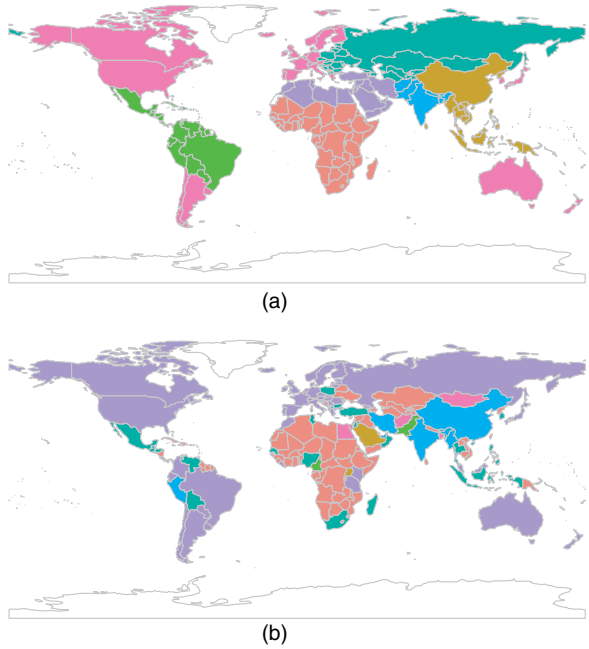

Fig. 2. (a) World Health Organization super-regions (the pink super-region corresponds to wealthy countries; the remaining regions are defined on the basis of geographic contiguity) and (b) super-regions found by clustering based on ground measurements of $PM_{2.5}$ concentration (countries for which we have no ground monitor measurements are coloured red)

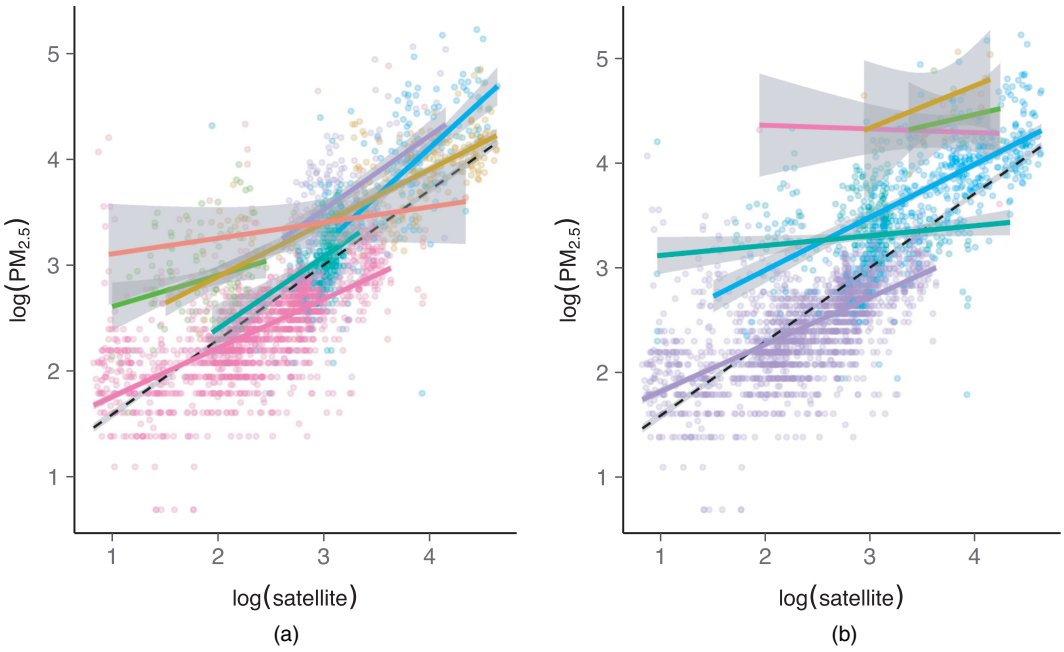

Fig. 3. Graphics in model building (here, evidence that a single linear trend is insufficient): (a) the same as Fig. 1(b), but also showing independent linear models fitted within each World Health Organization super-region; (b) the same as (a), but the linear models are fitted within each of the cluster regions shown in Fig. 2(b)

strategy to build a small network of three simple models for predicting $PM_{2.5}$ concentrations on a global scale.

The simplest predictive model that we can fit assumes that the satellite data product is a good predictor of the ground monitor data after a simple affine adjustment. In fact, this was the model that was used by the global burden of disease project before the 2016 update (Forouzanfar *et al.*, 2015). Fig. 1(b) shows a straight line that fits the data on a log–log-scale reasonably well ($R^2 \approx 0.6$). Discretization artefacts at the lower values of concentrations of $PM_{2.5}$ are also clearly visible.

To improve the model, we need to think about possible sources of heterogeneity. For example, we know that developed and developing countries have different levels of industrialization and hence different air pollution. We also know that desert sand can be a large source of $PM_{2.5}$. If these differences are not appropriately captured by the satellite data product, fitting only a single regression line could leave us in danger of falling prey to Simpson's paradox (that a trend can reverse when data are grouped).

To expand out our network of models, we consider two possible groupings of countries. The World Health Organization super-regions (Fig. 2(a)) separate out rich countries and divide the remaining countries into six geographically contiguous regions. These regions have not been constructed with air pollution in mind, so we also constructed a different division based on a six-component hierarchical clustering of ground monitor measurements of $PM_{2.5}$ (Fig. 2(b)). The seventh region constructed this way is the collection of all countries for which we do not have ground monitor data.

When the trends for each of these regions are plotted individually (Fig. 3), it is clear that some ecological bias would enter the analysis if we used only a single linear regression. We

also see that some regions, particularly sub-Saharan Africa (red in Fig. 3(a)) and clusters 1 and 6 (pink and yellow in Fig. 3(b)), do not have enough data to pin down the linear trend comprehensively. This suggests that some borrowing of strength through a multilevel model may be appropriate.

From this preliminary data analysis, we have constructed a network of three potential models. Model 1 is a simple linear regression. Model 2 is a multilevel model where observations are stratified by World Health Organization super-region. Model 3 is a multilevel model where observations are stratified by *clustered* super-region.

These three models will be sufficient for demonstrating our proposed workflow, but this is a smaller network of models than we would use for a comprehensive analysis of the $PM_{2.5}$ data. Shaddick *et al*. (2018), for example, also considered smaller regions, country level variation and a spatial model for the varying coefficients. Further calibration covariates can also be included.

3. Fake data can be almost as valuable as real data for building your model

The exploratory data analysis resulted in a network of three models: one linear regression model and two different linear multilevel models. To specify these models fully, we need to specify prior distributions on all the parameters. If we specify proper priors for all parameters in the model, a Bayesian model yields a joint prior distribution on parameters and data, and hence a prior marginal distribution for the data, i.e. Bayesian models with proper priors are *generative models*. The main idea in this section is that we can visualize simulations from the prior marginal distribution of the data to assess the consistency of the chosen priors with domain knowledge.

The main advantage to assessing priors based on the prior marginal distribution for the data is that it reflects the interplay between the prior distribution on the parameters and the likelihood. This is a vital component of understanding how prior distributions actually work for a given problem (Gelman *et al*., 2017). It also explicitly reflects the idea that we cannot fully understand the prior by fixing all except one parameter and assessing the effect of the unidimensional marginal prior. Instead, we need to assess the effect of the prior as a multivariate distribution.

The prior distribution over the data enables us to extend the concept of a weakly informative prior (Gelman *et al*., 2008) to be more aware of the role of the likelihood. In particular, we say that a prior leads to a *weakly informative joint prior data-generating process* if draws from the prior data-generating distribution $p(y)$ could represent any data set that could plausibly be observed. As with the standard concept of weakly informative priors, it is important that this prior predictive distribution for the data has at least some mass around extreme but plausible data sets. However, there should be no mass on completely implausible data sets. We recommend assessing how informative the prior distribution on the data is by generating a 'flip book' (a series of visualizations to scroll through) of simulated data sets that can be used to investigate the variability and multivariate structure of the distribution.

To demonstrate the power of this approach, we return to the multilevel model for the $PM_{2.5}$ data. Mathematically, the model will look like $y_{ij} \sim N\{\beta_0 + \beta_{0j} + (\beta_1 + \beta_{1j})x_{ij}, \sigma^2\}$, $\beta_{0j} \sim N(0, \tau_0^2)$, $\beta_{1j} \sim N(0, \tau_1^2)$, where y_{ij} is the logarithm of the observed concentration of $PM_{2.5}$, x_{ij} is the logarithm of the estimate from the satellite model, i ranges over the observations in each super-region, j ranges over the super-regions and σ, τ_0, τ_1, β_0 and β_1 need prior distributions.

Consider some priors of the sort that are sometimes recommended as being vague: $\beta_k \sim N(0, 100)$ and $\tau_k^2 \sim \text{Inv-gamma}(1, 100)$. The data that are generated by using these priors and shown in Fig. 4(a) are completely impossible for this application; note the y-axis limits and recall that the data are on the log-scale. This is primarily because the vague priors do not actually respect our contextual knowledge.

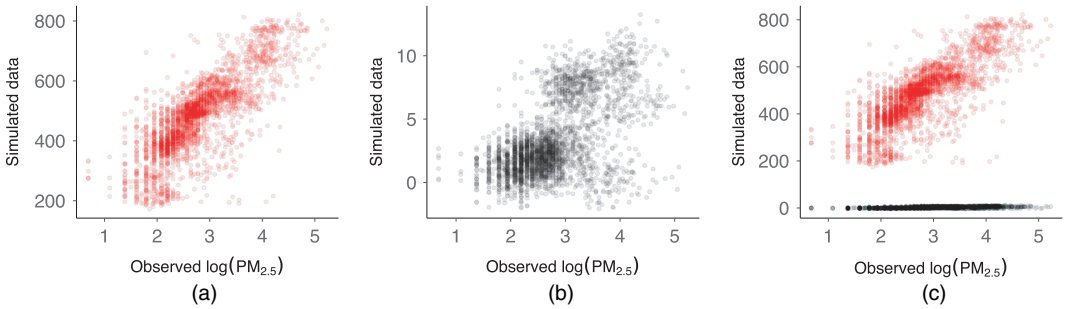

Fig. 4. Visualizing the prior predictive distribution: (a) and (b) show realizations from the prior predictive distribution using priors for the βs and τs that are vague and weakly informative respectively; the same $N_+(0, 1)$ prior is used for σ in both cases; simulated data are plotted on the y-axis and observed data on the x-axis; because the simulations under the vague and weakly informative priors are so different, the y-axis scales used in panels (a) and (b) also differ dramatically; (c) emphasizes the difference in the simulations by showing the red points from (a) and the black points from (b) plotted with the same y-axis

We know that the satellite estimates are reasonably faithful representations of the concentration of $PM_{2.5}$, so a more sensible set of priors would be centred near models with intercept 0 and slope 1. An example of this would be $\beta_0 \sim N(0, 1)$, $\beta_1 \sim N(1, 1)$ and $\tau_k \sim N_+(0, 1)$, where N_+ is the half-normal distribution. Data that are generated by this model are shown in Fig. 4(b). Although it is clear that this realization corresponds to quite a miscalibrated satellite model (especially when we remember that we are working on the log-scale), it is considerably more plausible than the model with vague priors.

We argue that the tighter priors are still only weakly informative, in that the implied data-generating process can still generate data that are much more extreme than we would expect from our domain knowledge. In fact, when repeating the simulation that is shown in Fig. 4(b) many times we found that the data that are generated by using these priors can produce data points with more than 22 000 μg m^{-3}, which is still a very high number in this context.

The prior predictive distribution is a powerful tool for understanding the structure of our model before we make a measurement, but its density evaluated at the measured data also plays the role of the marginal likelihood which is commonly used in model comparison. Unfortunately the utility of the prior predictive distribution to evaluate the model does not extend to utility in selecting between models. For further discussion see Gelman *et al*. (2017).

4. Graphical Markov chain Monte Carlo diagnostics: moving beyond trace plots

Constructing a network of models is only the first step in the Bayesian workflow. Our next job is to fit them. Once again, visualizations can be a key tool in doing this well. Traditionally, Markov chain Monte Carlo (MCMC) diagnostic plots consist of trace plots and auto-correlation functions. We find that these plots can be helpful to understand problems that have been caught by numerical summaries such as the potential scale reduction factor $\hat{R}$ (Stan Development Team (2017b), section 30.3), but they are not always needed as part of the workflow in the many settings where chains mix well.

For general MCMC methods it is difficult to do any better than between- and within-summary comparisons, following up with trace plots as needed. But, if we restrict our attention to HMC sampling and its variants, we can obtain much more detailed information about the performance of the Markov chain (Betancourt, 2017). We know that the success of HMC sampling requires that the geometry of the set containing the bulk of the posterior probability mass (which we

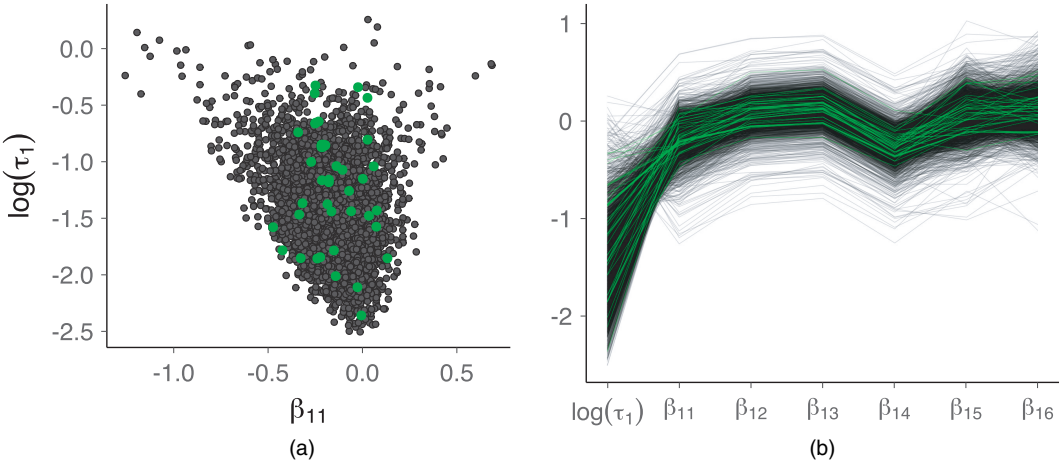

Fig. 5. Diagnostic plots for HMC sampling (models were fitted by using the RStan interface to Stan 2.17 (Stan Development Team, 2017a)): (a) for model 3, a bivariate plot of the log-standard-deviation of the cluster level slopes (y-axis) against the slope for the first cluster (x-axis) (the green dots indicate starting points of divergent transitions; this plot can be made by using mcmc_scatter in bayesplot); (b) for model 3, a parallel co-ordinates plot showing the cluster level slope parameters and their log-standard-deviation $\log(\tau_1)$ (the green lines indicate starting points of divergent transitions; this plot can be made by using mcmc_parcoord in bayesplot)

call the typical set) is fairly smooth. It is not possible to check this condition mathematically for most models, but it can be checked numerically. It turns out that, if the geometry of the typical set is non-smooth, the path taken by leapfrog integrator that defines the HMC proposal will rapidly diverge from the energy conserving trajectory.

Diagnosing divergent numerical trajectories precisely is difficult, but it is straightforward to identify these divergences heuristically by checking whether the error in the Hamiltonian crosses a large threshold. Occasionally this heuristic falsely flags stable trajectories as divergent, but we can identify these false positive results visually by checking whether the samples that are generated from divergent trajectories are distributed in the same way as the non-divergent trajectories. Combining this simple heuristic with visualization greatly increases its value.

Visually, a concentration of divergences in small neighbourhoods of parameter space, however, indicates a region of high curvature in the posterior that obstructs exploration. These neighbourhoods will also impede any MCMC method based on local information, but to our knowledge only HMC sampling has enough mathematical structure to be able to diagnose these features reliably. Hence, when we are using HMC sampling for our inference, we can use visualization to assess the convergence of the MCMC method and also to understand the geometry of the posterior.

There are several plots that we have found useful for diagnosing troublesome areas of the parameter space, in particular bivariate scatter plots that mark the divergent transitions (Fig. 5(a)) and parallel co-ordinate plots (Fig. 5(b)). These visualizations are sufficiently sensitive to differentiate between models with a non-smooth typical set and models where the heuristic has given a false positive result. This makes them an indispensable tool for understanding the behaviour of an HMC algorithm when applied to a particular target distribution.

If an HMC algorithm were struggling to fit model 3, the divergences would be clustered in the parameter space. Examining the bivariate scatter plots (Fig. 5(a)), there is no obvious pattern to the divergences. Similarly, the parallel co-ordinate plot (Fig. 5(b)) does not show any particular structure. This indicates that the divergences that are found are most probably false

positive results. For contrast, the on-line supplementary material contains the same plots for a model where HMC sampling fails to compute a reliable answer. In this case, the clustering of divergences is pronounced and the parallel co-ordinate plot clearly indicates that all the divergent trajectories have the same structure.

5. How did we do?: posterior predictive checks are vital for model evaluation

The idea behind posterior predictive checking is simple: if a model is a good fit we should be able to use it to generate data that resemble the data that we observed. This is similar in spirit to the prior checks that were considered in Section 3, except now we have a data-informed data-generating model. This means that we can be much more stringent in our comparisons. Ideally, we would compare the model predictions with an independent test data set, but this is not always feasible. However, we can still do some checking and predictive performance assessments by using the data that we already have.

To generate the data that are used for posterior predictive checks we simulate from the posterior predictive distribution $p(\tilde{y} \mid y) = \int p(\tilde{y} \mid \theta)\, p(\theta \mid y)\, \mathrm{d}\theta$, where y are our current data, $\tilde{y}$ are our

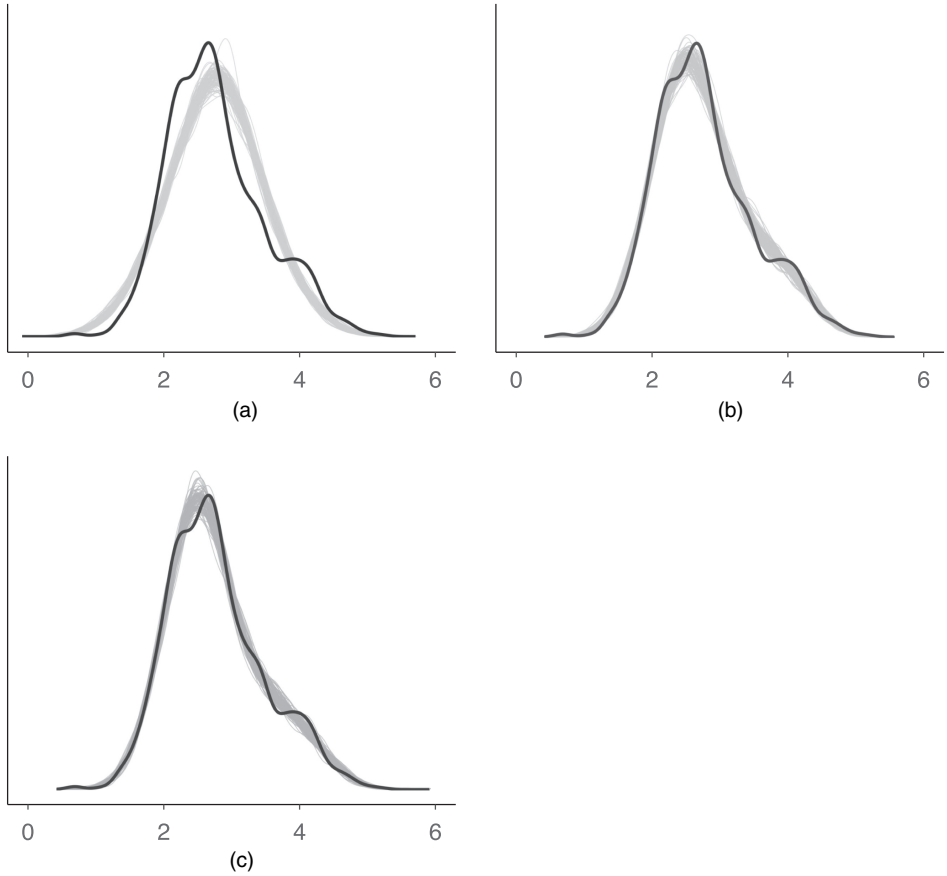

Fig. 6. Kernel density estimate of the observed data set y (dark curves), with density estimates for 100 simulated data sets y_{rep} drawn from the posterior predictive distribution (thin, lighter curves) (these plots can be produced using ppc_dens_overlay in the bayesplot package): (a) model 1; (b) model 2; (c) model 3

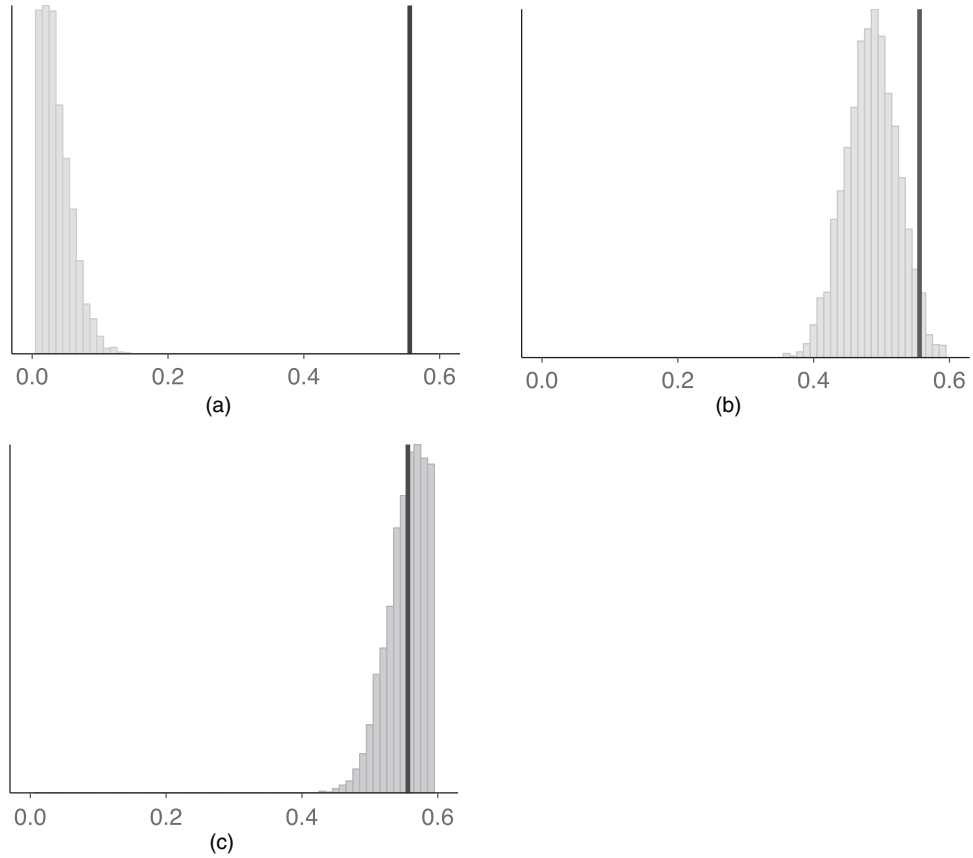

Fig. 7. Histograms of statistics skew(y_{rep}) computed from 4000 draws from the posterior predictive distribution (the dark vertical line is computed from the observed data; these plots can be produced using ppc_stat in the bayesplot package): (a) model 1; (b) model 2; (c) model 3

new data to be predicted and θ are our model parameters. Posterior predictive checking is mostly qualitative. By looking at some important features of the data and the replicated data, which were not explicitly included in the model, we may find a need to extend or modify the model.

For each of the three models, Fig. 6 shows the distributions of many replicated data sets drawn from the posterior predictive distribution (thin light curves) compared with the empirical distribution of the observed outcome (the thick dark curve). From these plots it is evident that the multilevel models (models 2 and 3) can simulate new data that are more similar to the observed log($PM_{2.5}$) values than the model without any hierarchical structure (model 1).

Posterior predictive checking makes use of the data twice: once for the fitting and once for the checking. Therefore it is a good idea to choose statistics that are orthogonal to the model parameters. If the test statistic is related to one of the model parameters, e.g. if the mean statistic is used for a Gaussian model with a location parameter, the posterior predictive checks may be less able to detect conflicts between the data and the model. Our running example uses a Gaussian model so in Fig. 7 we investigate how well the posterior predictive distribution captures skewness. Model 3, which used data-adapted regions, is best at capturing the observed skewness, whereas model 2 does a satisfactory job and the linear regression (model 1) totally fails.

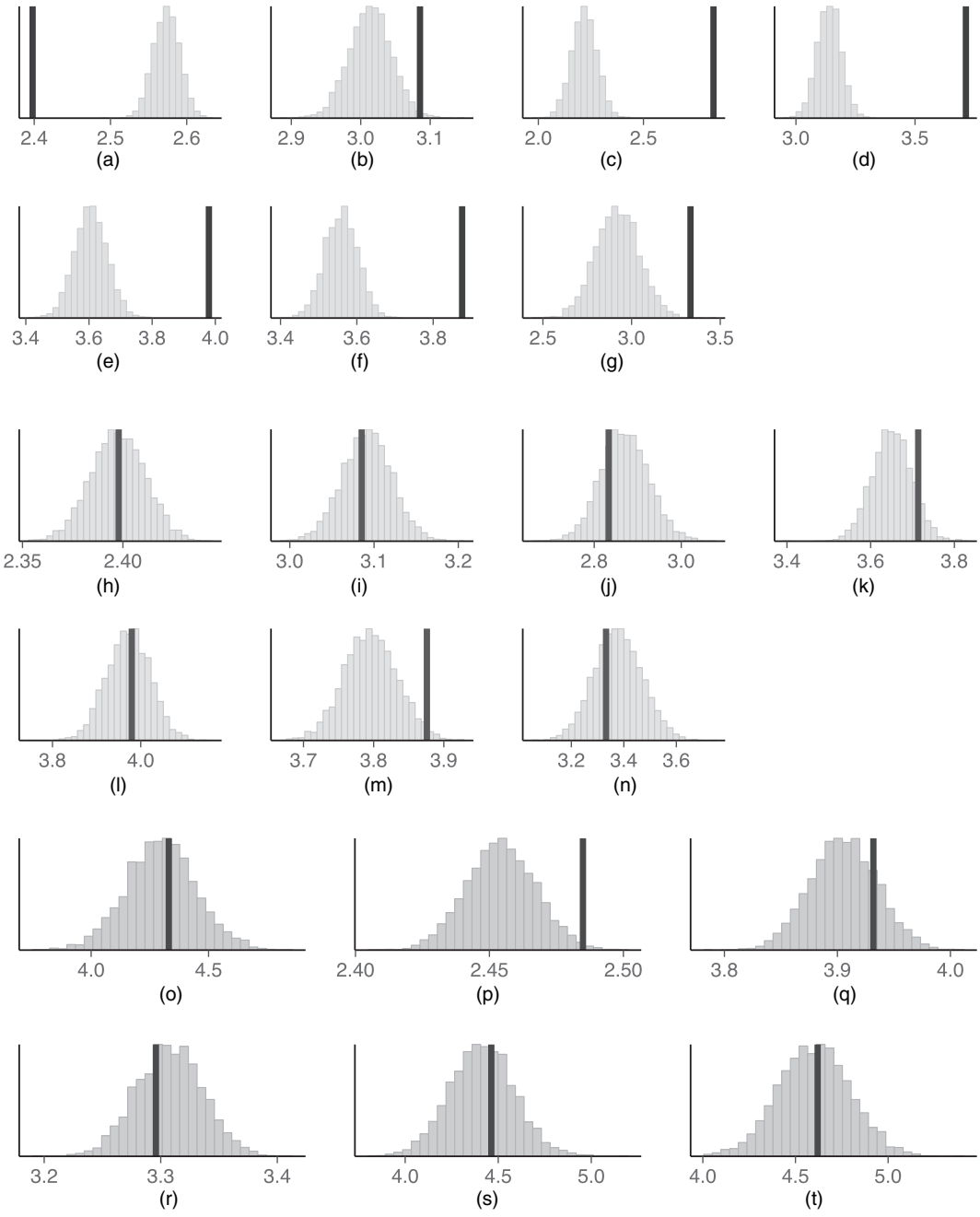

Fig. 8. Checking posterior predictive test statistics, in this case the medians, within region (the vertical lines are the observed medians; the facets are labelled by number in (o)–(t) because they represent groups found by the clustering algorithm rather than actual super-regions; these grouped plots can be made using ppc_stat_grouped in the bayesplot package): (a)–(g) model 1; (h)–(n) model 2; (o)–(t) model 3; (a), (h) high income super-region; (b), (i) eastern Europe–central Europe–central Asia; (c), (j) Latin America–Caribbean; (d), (k) north Africa–Middle East; (e), (l) south Asia; (f), (m) south-east Asia–east Asia–Oceania; (g), (n), sub-Saharan Africa; (o) group 1; (p) group 2; (q) group 3; (r) group 4; (s) group 5; (t) group 6

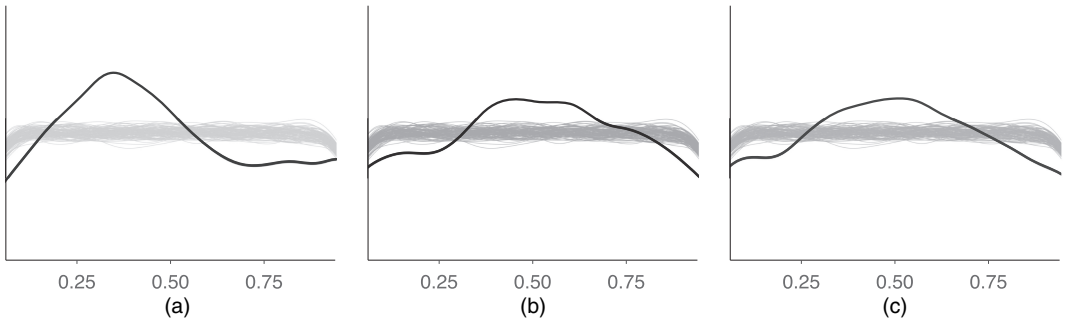

Fig. 9. Graphical check of the LOO cross-validated probability integral transform (——, simulations from the standard uniform distribution; ——, density of the computed LOO probability integral transforms) (similar plots can be made using ppc_dens_overlay and ppc_loo_pit in the bayesplot package; the downward slope near 0 and 1 on the 'uniform' histograms is an edge effect due to the density estimator used and can be safely discounted): (a) model 1; (b) model 2; (c) model 3

We can also perform similar checks within levels of a grouping variable. For example, in Fig. 8 we split both the outcome and the posterior predictive distribution according to region and check the median values. The two hierarchical models give a better fit to the data at the group level, which in this case is unsurprising.

In cross-validation, double use of data is partially avoided and test statistics can be better calibrated. When performing leave-one-out (LOO) cross-validation we usually work with univariate posterior predictive distributions, and thus we cannot examine properties of the joint predictive distribution. To check specifically that predictions are calibrated, the usual test is to look at the LOO cross-validation predictive cumulative density function values, which are asymptotically uniform (for continuous data) if the model is calibrated (Gelfand *et al*., 1992; Gelman *et al*., 2013).

The plots that are shown in Fig. 9 compare the density of the computed LOO probability integral transforms (the thick dark curve) *versus* 100 simulated data sets from a standard uniform distribution (the thin light curves). We can see that, although there is some clear miscalibration in all cases, the hierarchical models are an improvement over the single-level model.

The shape of the miscalibration in Fig. 9 is also meaningful. The frown shapes that are exhibited by models 2 and 3 indicate that the univariate predictive distributions are too broad compared with the data, which suggests that further modelling will be necessary to reflect the uncertainty accurately. One possibility would be to subdivide the super-regions further to capture within-region variability better (Shaddick *et al*., 2018).

6. Pointwise plots for predictive model comparison

Visual posterior predictive checks are also useful for identifying unusual points in the data. Unusual data points come in two flavours: outliers and points with high leverage. In this section, we show that visualization can be useful for identifying both types of data point. Examining these unusual observations is a critical part of any statistical workflow, as these observations give hints about how the model may need to be modified. For example, they may indicate that the model should use non-linear instead of linear regression, or that the observation error should be modelled with a heavier-tailed distribution.

The main tool in this section is the one-dimensional cross-validated LOO predictive distribution $p(y_i \mid y_{-i})$. Gelfand *et al*. (1992) suggested examining the LOO log-predictive density values

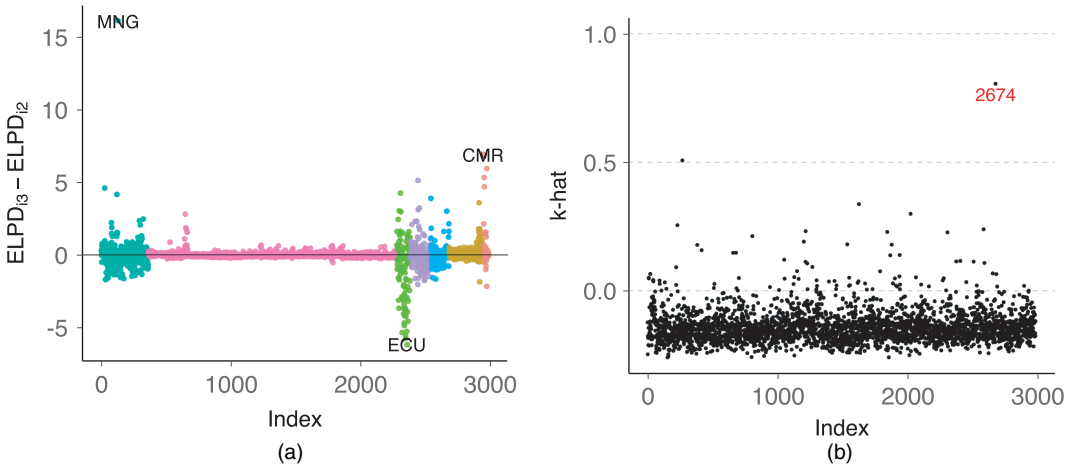

Fig. 10. Model comparisons by using LOO cross-validation: (a) the difference in pointwise values obtained from LOO PSIS for model 3 compared with model 2 coloured by World Health Organization cluster (see Fig. 1(b) for the key; positive values indicate that model 3 outperformed model 2); (b) $\hat{k}$-diagnostics from LOO PSIS for model 2 (the 2674th data point (the only data point from Mongolia) is highlighted by the $\hat{k}$-diagnostic as being influential on the posterior)

(they called them conditional predictive ordinates) to find observations that are difficult to predict. This idea can be extended to model comparison by looking at which model best captures each left-out data point. Fig. 10(a) shows the difference between the expected log-predictive densities ELPD for the individual data points estimated by using Pareto-smoothed importance sampling (PSIS) (Vehtari *et al.*, 2017a,b). Model 3 appears to be slightly better than model 2, especially for difficult observations like the station in Mongolia.

In addition to looking at the individual LOO log-predictive densities, it is useful to look at how influential each observation is. Some of the data points may be difficult to predict but not necessarily influential, i.e. the predictive distribution does not change much when they are left out. One way to look at the influence is to look at the difference between the full data log-posterior predictive density and the LOO log-predictive density.

We recommend computing the LOO log-predictive densities by using the PSIS LOO method as implemented in the `loo` package (Vehtari *et al.*, 2017c). A key advantage of using the PSIS LOO method to compute the LOO densities is that it automatically computes an empirical estimate of how similar the full data predictive distribution is to the LOO predictive distribution for each left-out point. Specifically, it computes an empirical estimate $\hat{k}$ of $k = \inf\{k' > 0 : D_{1/k'}(p||q) < \infty\}$, where

$$D_\alpha(p||q) = \frac{1}{\alpha - 1} \log \left\{ \int_\Theta p(\theta)^\alpha q(\theta)^{1-\alpha} \mathrm{d}\theta \right\}$$

is the α-Rényi divergence (Yao *et al.*, 2018). If the jth LOO predictive distribution has a large $\hat{k}$-value when used as a proposal distribution for the full data predictive distribution, it suggests that y_j is a highly influential observation.

Fig. 10(b) shows the $\hat{k}$-diagnostics from the PSIS LOO method for our model 2. The 2674th data point is highlighted by the $\hat{k}$-diagnostic as being influential on the posterior. If we examine the data we find that this point is the only observation from Mongolia and corresponds to a measurement $(x, y) = (\log(\text{satellite}), \log(\text{PM}_{2.5})) = (1.95, 4.32)$, which would look like an outlier if highlighted in the scatter plot in Fig. 1(b). By contrast, under model 3 the $\hat{k}$-value for the

Mongolian observation is significantly lower ($\hat{k} \approx 0.5$) indicating that that point is better resolved in model 3.

7. Discussion

Visualization is probably the most important tool in an applied statistician's toolbox and is an important complement to quantitative statistical procedures (Buja *et al*., 2009). In this paper, we have demonstrated that it can be used as part of a strategy to compare models, to identify ways in which a model fails to fit, to check how well our computational methods have resolved the model, to understand the model sufficiently well to be able to set priors and to improve the model iteratively.

The last of these tasks is a little controversial as using the measured data to guide model building raises the concern that the resulting model will generalize poorly to new data sets. A different objection to using the data twice (or even more) comes from ideas around hypothesis testing and unbiased estimation, but we are of the opinion that the danger of overfitting the data is much more concerning (Gelman and Loken, 2014).

In the visual workflow that we have outlined in this paper, we have used the data to improve the model in two places. In Section 3 we proposed prior predictive checks with the recommendation that the data-generating mechanism should be broader than the distribution of the observed data in line with the principle of weakly informative priors. In Section 5 we recommended undertaking careful calibration checks as well as checks based on summary statistics, and then updating the model accordingly to cover the deficiencies that are exposed by this procedure. In both of these cases, we have made recommendations that aim to reduce the danger. For the prior predictive checks, we recommend aiming for a prior data-generating process that can produce plausible data sets, not necessarily data sets that are indistinguishable from observed data. For the posterior predictive checks, we ameliorate the concerns by checking carefully for influential measurements and proposing that model extensions be weakly informative extensions that are still centred on the previous model (Simpson *et al*., 2017).

Regardless of concerns that we have about using the data twice, the workflow that we have described in this paper (perhaps without the stringent prior and posterior predictive checks) is common in applied statistics. As academic statisticians, we have a duty to understand the consequences of this workflow and offer concrete suggestions to make the practice of applied statistics more robust.

Acknowledegements

The authors thank Gavin Shaddick and Matthew Thomas for their help with the $PM_{2.5}$ example, Ari Hartikainen for suggesting the parallel co-ordinates plot, Ghazal Fazelnia for finding an error in our map of ground monitor locations, Eren Metin Elçi for alerting us to a discrepancy between our text and code, and the Sloan Foundation, Columbia University, US National Science Foundation, Institute for Education Sciences, Office of Naval Research and Defense Advanced Research Projects Agency for financial support.

References

Betancourt, M. (2017) A conceptual introduction to Hamiltonian Monte Carlo. *Preprint arXiv:1701.02434*. Columbia University, New York.

Buja, A., Cook, D., Hofmann, H., Lawrence, M., Lee, E.-K., Swayne, D. F. and Wickham, H. (2009) Statistical inference for exploratory data analysis and model diagnostics. *Phil. Trans. R. Soc. Lond.*, **367**, 4361–4383.

Forouzanfar, M. H., Alexander, L., Anderson, H. R., Bachman, V. F., Biryukov, S., Brauer, M., Burnett, R., Casey, D., Coates, M. M., Cohen, A. *et al.* (2015) Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet*, **386**, 2287–2323.

Gabry, J. (2017) bayesplot: plotting for Bayesian models. *R Package Version 1.3.0.* Columbia University, New York. (Available from `http://mc-stan.org/bayesplot`.)

Gelfand, A. E., Dey, D. K. and Chang, H. (1992) Model determination using predictive distributions with implementation via sampling-based methods (with discussion). In *Bayesian Statistics 4* (eds. J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith), pp. 147–167. Oxford: Clarendon.

Gelman, A. (2004) Exploratory data analysis for complex models. *J. Computnl Graph. Statist.*, **13**, 755–779.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2013) Marginal predictive checks. In *Bayesian Data Analysis*, 3rd edn, ch. 6. Boca Raton: Chapman and Hall–CRC.

Gelman, A., Jakulin, A., Pittau, M. G. and Su, Y.-S. (2008) A weakly informative default prior distribution for logistic and other regression models. *Ann. Appl. Statist.*, **2**, 1360–1383.

Gelman, A. and Loken, E. (2014) The statistical crisis in science: data-dependent analysis—a "garden of forking paths"—explains why many statistically significant comparisons don't hold up. *Am. Scient.*, **102**, no. 6, 460.

Gelman, A., Simpson, D. and Betancourt, M. (2017) The prior can often only be understood in the context of the likelihood. *Entropy*, **19**, no. 10, article 555.

R Core Team (2017) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Shaddick, G., Thomas, M. L., Green, A., Brauer, M., van Donkelaar, A., Burnett, R., Chang, H. H., Cohen, A., Van Dingenen, R., Dora, C., Gumy, S., Liu, Y., Martin, R., Waller, L. A., West, J., Zidek, J. V. and Prüss-Ustün, A. (2018) Data integration model for air quality: a hierarchical approach to the global estimation of exposures to ambient air pollution. *Appl. Statist.*, **68**, 231–253.

Simpson, D., Rue, H., Riebler, A., Martins, T. G. and Sørbye, S. H. (2017) Penalising model component complexity: a principled, practical approach to constructing priors. *Statist. Sci.*, **32**, 1–28.

Stan Development Team (2017a) RStan: the R interface to Stan, version 2.16.1. Stan Development Team. (Available from `http://mc-stan.org`.)

Stan Development Team (2017b) *Stan Modeling Language User's Guide and Reference Manual*, *Version 2.16.0*. Stan Development Team. (Available from `http://mc-stan.org`.)

Vehtari, A., Gelman, A. and Gabry, J. (2017a) Pareto smoothed importance sampling. *Preprint arXiv:1507.02646*. Aalto University, Espoo.

Vehtari, A., Gelman, A. and Gabry, J. (2017b) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statist. Comput.*, **27**, 1413–1432.

Vehtari, A., Gelman, A. and Gabry, J. (2017c) loo: efficient leave-one-out cross-validation and WAIC for Bayesian models. *R Package Version 1.0.0.* (Available from `http://mc-stan.org/loo`.) Aalto University, Espoo.

Wickham, H. (2009) *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Yao, Y., Vehtari, A., Simpson, D. and Gelman, A. (2018) Yes, but did it work?: Evaluating variational inference. *Proc. Mach. Learn. Res.*, **80**, 5581–5590.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material: Visualization in Bayesian workflow'.