International Institute for
Applied Systems Analysis
Schlossplatz 1
A-2361 Laxenburg, Austria

Tel: +43 2236 807 342
Fax: +43 2236 71313
E-mail: repository@iiasa.ac.at
Web: www.iiasa.ac.at

Working Paper

Cross-comparison of Integration Methods in the Case Study of Sea Level Pressure Model Ensemble

Anna Shchiptsova (shchipts@iiasa.ac.at)
Benjamin Renard (benjamin.renard@irstea.fr)
Elena Rovenskaya (rovenska@iiasa.ac.at)

Approved by

Elena Rovenskaya
Program Director, Advanced Systems Analysis
September 2016

ZVR 524808900

Contents

Abstract

The laws that drive a complex social-environmental system are never perfectly understood. Due to the high complexity of the underlying system processes, researches tend to create an ensemble of multiple models, which describe the studied phenomenon using different modeling approaches and primary assumptions. A set of ensemble outcomes (usually represented by a family of probability distributions) then needs to be integrated into one estimate in order to install an ensemble into the modelling chain or provide support for the informed decision making. This paper deals with the case study of the sea level pressure model ensemble. We examine performance of the two alternative integration methods, the posterior integration method and the Bayesian Information Criterion (BIC) method. In the latter case, the integration approach is data-driven, and subsequently, we introduce a cross validation procedure to compare the resultant integrated estimates of the sea level pressure.

Acknowledgments

The authors would like to acknowledge DG research for funding through the FP7-funded COMPLEX project #308601, www.complex.ac.uk.

About the Authors

Anna Shchiptsova is a Research Scholar with Advanced Systems Analysis Program at IIASA.

Benjamin Renard is a Researcher in Stochastic Hydrology at Institut de recherche pour l'ingénierie de l'agriculture et de l'environnement (IRSTEA), France.

Elena Rovenskaya is the Director of the Advanced Systems Analysis Program at IIASA; she is also a Research Scholar at the Faculty of Computational Mathematics and Cybernetics, Lomonosov Moscow State University, Russia.

Cross-comparison of Integration Methods in the Case Study of Sea Level Pressure Model Ensemble

Anna Shchiptsova
Benjamin Renard
Elena Rovenskaya

1 Introduction

The laws that drive a complex social-environmental system are never perfectly understood (Kryazhimskiy, 2014). In general, a single model cannot capture the complexity of the underlying system processes. As a result, researchers tend to create a suite of models, which are based on different primary assumptions about the studied phenomenon. Each model usually enters an ensemble with its intrinsic uncertainties. Namely, the choices made in individual model design often imply model uncertainties in initial conditions, uncertainties in boundary conditions, parameter and structural uncertainties (Tebaldi and Knutti, 2007). In the end, we deal with a set of statistically imprecise ensemble outcomes.

However, if we want to install an ensemble into a modeling chain or to provide support for the informed decision making, models' output needs to be integrated into one estimate. Here, we make an assumption that if individual models are independent, their errors might at least partly cancel after integration, resulting in a multi-model average that is more skillful than its constitutive terms (Tebaldi and Knutti 2007). Several approaches to determine a combined estimate from ensemble members have been proposed in the literature. Reviews of existing integration methods can be found in Clemen (1989), which considers contributions from the forecasting, psychology, statistics, and management science literatures; Genest and Zidek (1986), which describe formal, often statistical, approaches to aggregation of a number of expert opinions (models); Tebaldi and Knutti (2007), which specifically outline methods used in climate modeling. In general, we can distinguish data-driven approaches to model integration, where individual estimates are combined into a weighted combination based on the models' performance in the past and the present. For example, this category of methods includes Bayesian Model Averaging (Hoeting et al., 1999) and methods of models weighting using information criteria (Burnham and Anderson, 2004). As opposed to data-driven techniques, Kryazhimskiy (2013, 2016) suggests the posterior integration methodology based on construction of posterior event in the product of the probability spaces associated with the prior model estimates.

This paper compares alternative integration methods in the case study of the sea level pressure model ensemble. The prior estimates on the sea level pressure are obtained within the COMPLEX project case studies (Renard et al., 2014; http://owsgip.itc.utwente.nl/projects/complex/). The suite of integration techniques contains the posterior integration method (Kryazhimskiy 2013, 2016) and the BIC method based on model weighting (Burnham and Anderson, 2004; Renard et al., 2014). In the latter case, the integration approach is data-driven. In this connection, we introduce a cross validation procedure to compare alternative integrated estimates. The rest of the paper is organized as follows. In section 2 we outline methodology for the cross-comparison of integration methods. Section 3 describes the case study of the sea level pressure model ensemble. Results of cross-comparison are presented and discussed in section 4.

2 Methodology

2.1 Integration Methods

We consider an ensemble of n sea level pressure models. By assumption, each model rounds the sea level pressure value off to some unit; subsequently, it gives an independent statistically imprecise estimate of the rounding unit (interval), where the true sea level pressure falls. Let z_0 be an actual unknown interval of the sea level pressure. We suppose that z_0 belongs to a non-empty finite set Z of all rounding intervals, whose number of elements is bigger than one. Thus, each model i represents the unknown z_0 as a probability p_i on Z. In addition, a probability p_{data} on Z is obtained from the series of instrumental observations on the sea level pressure.

In this research, we combine information from the model ensemble using two alternative integration methods. The posterior integration method (Kryazhimskiy, 2013; Kryazhimskiy, 2016) is based on the assumption that model outcomes are mutually compatible, i.e., we should observe identical outcomes after the use of a model ensemble. Formally, the probability distribution of prior estimates is

$$\pi(z) = \frac{p_1(z) * p_2(z) * \dots * p_n(z)}{\sum_{z' \in Z} p_1(z') * p_2(z') * \dots * p_n(z')}, \quad z \in Z \tag{1}$$

The integration operation possesses the algebraic properties of multiplication (Kryazhimskiy, 2013). In what follows, we call distribution (1) the *product* probability distribution.

Secondly, we consider the weighting of models in an ensemble based on the Bayesian Information Criterion (BIC)

$$\pi(z) = \sum_{i=1}^{n} \frac{e^{-BIC_i/2}}{\sum_{j=1}^{n} e^{-BIC_j/2}} * p_i(z), \quad z \in Z \tag{2}$$

The general rationale behind BIC weights calculation is given in Burnham and Anderson (2004). Renard et al. (2014) adapted the BIC method to the context of integrating the outcomes from several climate models (see COMPLEX project case studies). Note that the BIC approach is data-driven as weights computation utilizes information on prior estimates performance relative to the already observed measurements of the sea level pressure.

2.2 Performance Metrics

In general, our goal is to assess the performance of different ways of integration relative to the observed p_{data}. For this purpose, we compare means of probability distributions as follows

$$\Delta_{\mu}(\pi_i, p_{data}) = \frac{\mu_{(\pi_i)} - \mu_{(p_{data})}}{\sigma_{(p_{data})}} \tag{3}$$

where $\mu_{(\pi_i)}$ is the expected value of the integrated distribution π_i $(i = 1, 2)$, $\mu_{(p_{data})}$ is the expected value of the observed distribution p_{data} and $\sigma_{(p_{data})}$ is the standard deviation of the observed distribution p_{data}. Thus, we rate a deviation of the integrated mean from the observed one relative to the variability in observations. We call $\Delta_{\mu}(\pi_i, p_{data})$ a relative mean difference.

On the other hand, we estimate whether an integrated distribution replicates variability of the instrumental observations. In the first place, we measure variation in terms of standard deviation

$$\Delta_{\sigma}(\pi_i, p_{data}) = \frac{\sigma_{(\pi_i)} - \sigma_{(p_{data})}}{\sigma_{(p_{data})}} \tag{4}$$

where $\sigma_{(\pi_i)}$ is the standard deviation of the integrated distribution π_i $(i = 1, 2)$ and $\sigma_{(p_{data})}$ is the standard deviation of the observed distribution p_{data}. We call $\Delta_{\sigma}(\pi_i, p_{data})$ a relative difference in standard deviations.

Additionally, we compare difference in the shape of probability distributions using statistical distance (total variation distance)

$$\delta(\pi_i, p_{data}) = \frac{1}{2}\sum_{z \in Z} |\pi_i(z) - p_{data}(z)| \tag{5}$$

2.3 Cross Validation

We consider k-fold cross validation (Olson and Delen, 2008; Hastie et al., 2001) to study performance of the proposed integration methods. Let us assume that the number of available instrumental observations equals p. This dataset is divided into k subsets with equal (or near equal) number of points. Each of these subsets is used to estimate the performance of integration methods in separate cross validation runs. In each cross validation run, we calculate BIC weights using the remaining $k-1$ subsets of instrumental observations. We call a union of $k-1$ subsets used in the integration estimation a training dataset, and the k-th subset is called a validation dataset.

Overall, the cross validation process is repeated k times. Suppose that $p'_{data}(j)$ on Z is a probability obtained from instrumental observations in the training dataset, and $p''_{data}(j)$ on Z is a probability obtained from the validation dataset in cross validation run j. At first, we independently estimate performance of the integration methods in each run for the training and validation datasets applying formulas (3)-(5). After that, we determine average values of $\Delta_\mu(\pi_i, p'_{data}(j))$, $\Delta_\sigma(\pi_i, p'_{data}(j))$, $\delta(\pi_i, p'_{data}(j))$ for the training dataset, and average values of $\Delta_\mu(\pi_i, p''_{data}(j))$, $\Delta_\sigma(\pi_i, p''_{data}(j))$, $\delta(\pi_i, p''_{data}(j))$ for the validation dataset ($j = 1 \dots k$), where π_i is either the product probability distribution in (1) or the BIC probability distribution in (2).

Finally, we examine consistency of the performance metrics between the training and validation datasets for an integration method π_i $(i = 1, 2)$. For this purpose, we measure the root mean squared error between them

$$m(\pi_i) = \sqrt{\frac{1}{k}\sum_{j=1}^{k}(s(\pi_i, p'_{data}(j)) - s(\pi_i, p''_{data}(j)))^2} \qquad (6)$$

where s is a performance metric (Δ_μ, Δ_σ, δ).

3 Data

The data on the sea level pressure is extracted from the 20CR reanalysis (Compo et al., 2011) at longitude 5W, latitude 44N (near Marseille, France) for the period from 1871 to 2004. In each year, we take an average of daily observations per season. Namely, we aggregate values in the 3-month periods: January, February and March (season JFM); April, March and June (season AMJ), July, August and September (season JAS) and October, November and December (season OND). In total, each seasonal dataset comprises 134 points. On the other hand, initial data includes four seasonal sea level pressure time series, simulated by models 1 – 4 for the same period and region.

Here, we use 3-fold cross validation. That is, we divide period 1871 – 2004 into periods: 1871-1915 (number of data points 45), 1916 – 1959 (number of data points 44) and 1960 – 2004 (number of data points 45). Two parts of the time series are included in the training dataset and the remaining part constitutes the validation dataset in each cross validation run. Table 1 illustrates time periods corresponding to the 3-fold cross validation in the case study.

Table 1. Cross validation runs

	Years in the training dataset	Years in the validation dataset
Run 1	1871 – 1915, 1916 – 1959	1960 – 2004
Run 2	1871 – 1915, 1960 – 2004	1916 – 1959
Run 3	1916 – 1959, 1960 – 2004	1871 – 1915

We represent the observed seasonal sea level pressure time series and the sea level pressure time series simulated by models 1 – 4 in the training dataset as probability distributions. For this purpose, we put a uniform grid on the sea level pressure axis. The step size of the grid is defined as an average Scott's step size taken over five samples, which include observed and model time series. For a separate sample of values, the step size is calculated by the formula (Scott, 1979)

$$h = 3.5\hat{\sigma}/n^{1/3} \qquad (7)$$

where $\hat{\sigma}$ is the sample standard deviation and n is the number of sample values.

For each of the five seasonal sea level pressure time series we define the (empirical) probability of each grid cell to be the relative frequency of the sea level pressure values in the time series, which fall into that grid cell. In result, for each season and for each cross validation run we construct the data-based seasonal sea level pressure probability distribution p_{data} and seasonal sea level pressure probability distributions p_1, p_2, p_3, p_4 in the ensemble of models 1 – 4. After that, we compute performance metrics of the prior distributions p_1, p_2, p_3, p_4 relative to the observed distribution p_{data} in each cross validation run. Tables 2 – 9 present histograms for the resultant probability distributions and the values of performance metrics.

Table 2. Data-based probability distribution and probability distributions constructed from models 1 – 4 per cross validation run. Season JFM, grid step size 229.87 Pa

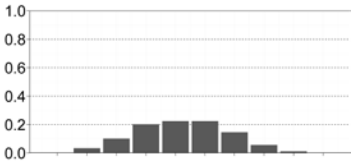
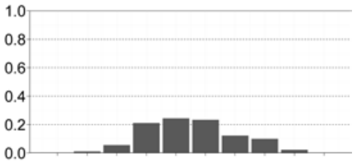
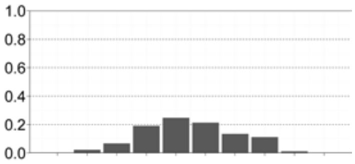
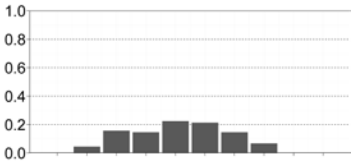
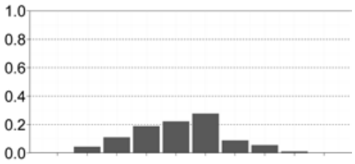
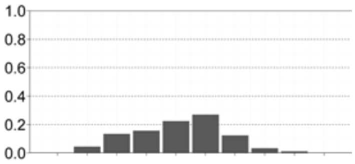
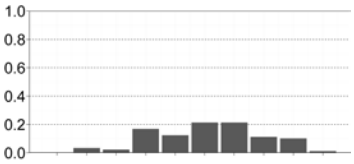
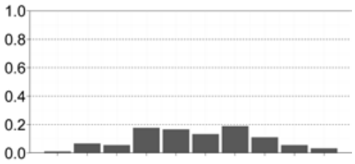
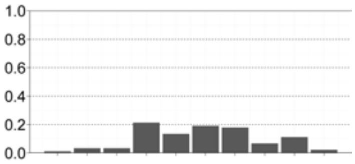
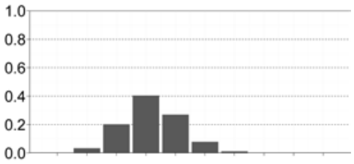
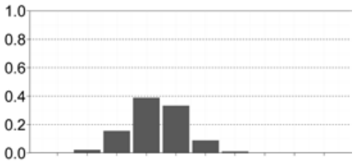
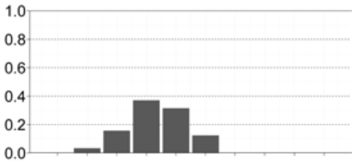
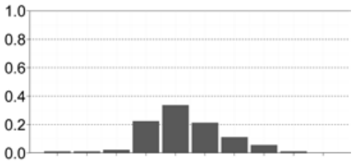
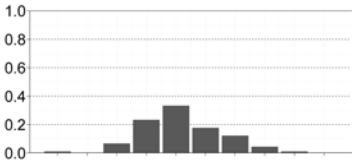
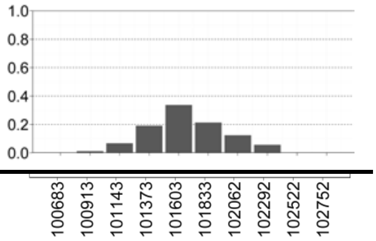
Distribution	Run 1	Run 2	Run 3
Data (training dataset)			
Model 1			
Model 2			
Model 3			
Model 4			

Table 3. Performance metrics (3) – (5) of probability distributions constructed from models 1 – 4 in the training period per cross validation run. Season JFM, grid step size 229.87 Pa

Distribution	Run 1	Run 2	Run 3
		Mean	
Data (training dataset)	0% (abs. 101654 Pa)	0% (abs. 101720 Pa)	0% (abs. 101711 Pa)
Model 1	-7%	-26%	-24%
Model 2	59%	13%	27%
Model 3	-67%	-78%	-73%
Model 4	5%	-21%	-13%
		Standard deviation	
Data (training dataset)	0% (abs. 352.68 Pa)	0% (abs. 344.80 Pa)	0% (abs. 354.93 Pa)
Model 1	5%	2%	-1%
Model 2	18%	38%	28%
Model 3	-35%	-36%	-36%
Model 4	-11%	-8%	-15%
		Statistical distance	
Data (training dataset)	0.000	0.000	0.000
Model 1	0.079	0.133	0.146
Model 2	0.225	0.211	0.213
Model 3	0.348	0.378	0.348
Model 4	0.146	0.133	0.090

Table 4. Data-based probability distribution and probability distributions constructed from models 1 – 4 per cross validation run. Season AMJ, grid step size 103.4 Pa

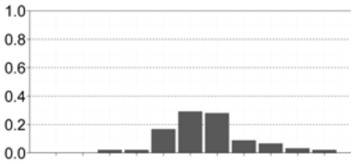
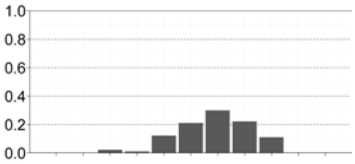
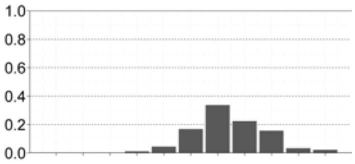
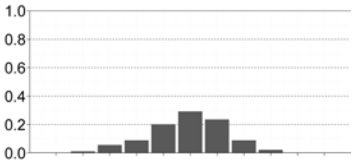
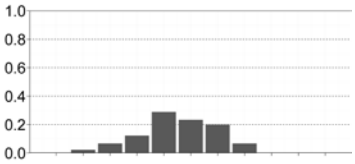
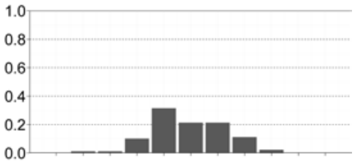
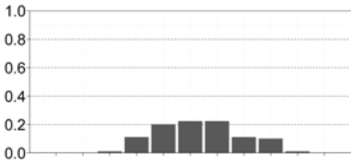
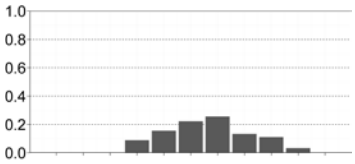
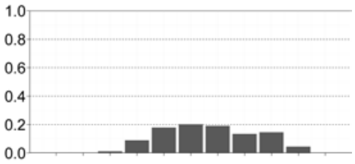
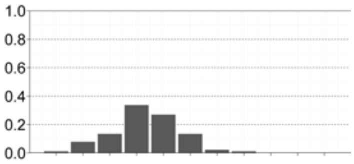
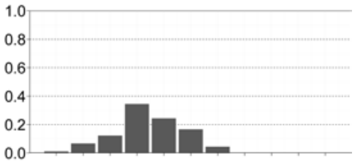
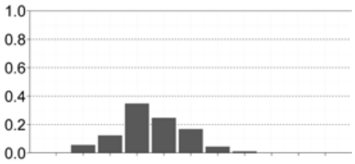
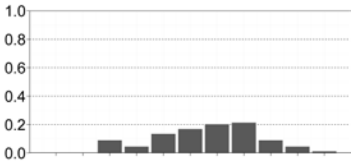
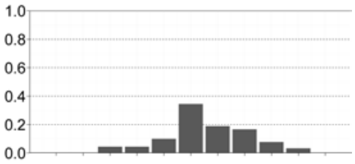
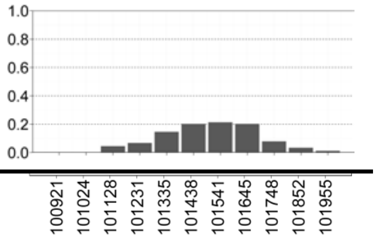
Distribution	Run 1	Run 2	Run 3
Data (training dataset)			
Model 1			
Model 2			
Model 3			
Model 4			

Table 5. Performance metrics (3) – (5) of probability distributions constructed from models 1 – 4 in the training period per cross validation run. Season AMJ, grid step size 103.4 Pa

Distribution	Run 1	Run 2	Run 3
		Mean	
Data (training dataset)	0% (abs. 101503 Pa)	0% (abs. 101528 Pa)	0% (abs. 101587 Pa)
Model 1	-48%	-101%	-115%
Model 2	-19%	-16%	-57%
Model 3	-148%	-182%	-218%
Model 4	1%	-22%	-61%
		Standard deviation	
Data (training dataset)	0% (abs. 160.63 Pa)	0% (abs. 138.90 Pa)	0% (abs. 137.73 Pa)
Model 1	-8%	4%	2%
Model 2	0%	15%	29%
Model 3	-18%	-5%	-6%
Model 4	22%	16%	29%
		Statistical distance	
Data (training dataset)	0.000	0.000	0.000
Model 1	0.146	0.367	0.427
Model 2	0.180	0.156	0.270
Model 3	0.618	0.633	0.719
Model 4	0.247	0.222	0.236

Table 6. Data-based probability distribution and probability distributions constructed from models 1 – 4 per cross validation run. Season JAS, grid step size 70.94 Pa

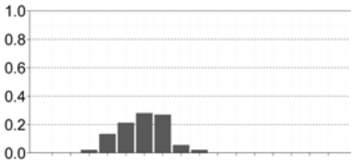
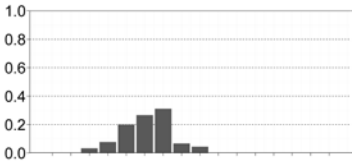
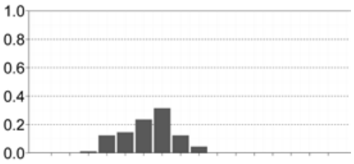
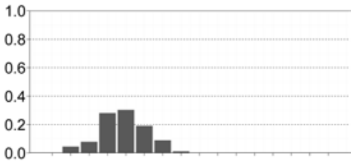
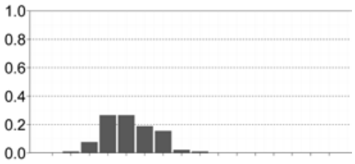
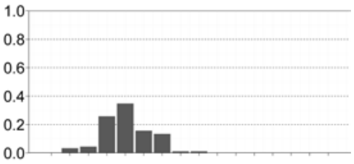
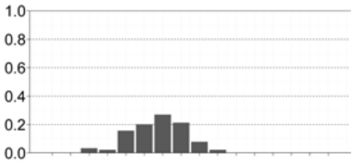
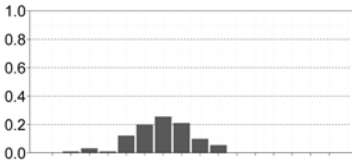
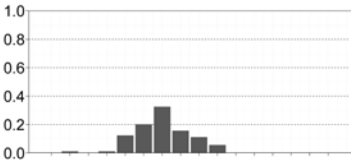
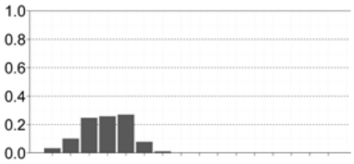
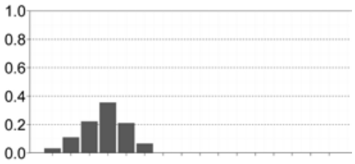
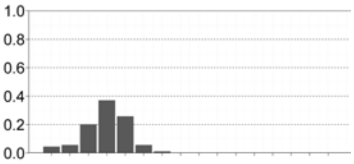
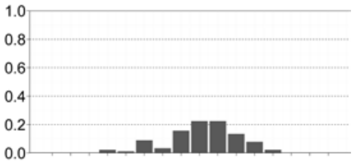
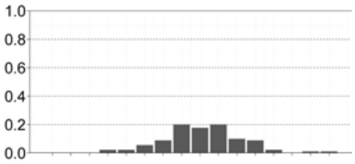
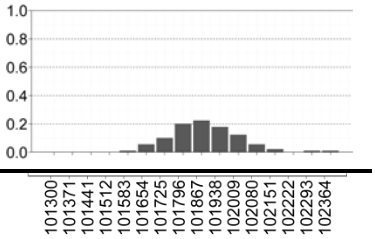
Distribution	Run 1	Run 2	Run 3
Data (training dataset)			
Model 1			
Model 2			
Model 3			
Model 4			

Table 7. Performance metrics (3) – (5) of probability distributions constructed from models 1 – 4 in the training period per cross validation run. Season JAS, grid step size 70.94 Pa

Distribution	Run 1	Run 2	Run 3
		Mean	
Data (training dataset)	0% (abs. 101647 Pa)	0% (abs. 101663 Pa)	0% (abs. 101673 Pa)
Model 1	-83%	-73%	-88%
Model 2	64%	59%	54%
Model 3	-155%	-174%	-169%
Model 4	255%	228%	215%
		Standard deviation	
Data (training dataset)	0% (abs. 90.89 Pa)	0% (abs. 94.48 Pa)	0% (abs. 97.33 Pa)
Model 1	-1%	2%	-5%
Model 2	16%	23%	8%
Model 3	0%	-11%	-13%
Model 4	49%	62%	41%
		Statistical distance	
Data (training dataset)	0.000	0.000	0.000
Model 1	0.337	0.311	0.404
Model 2	0.247	0.267	0.180
Model 3	0.539	0.622	0.652
Model 4	0.764	0.700	0.663

Table 8. Data-based probability distribution and probability distributions constructed from models 1 – 4 per cross validation run. Season OND, grid step size 162.89 Pa

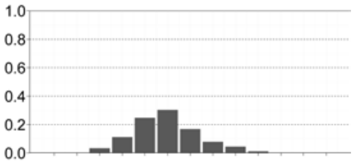
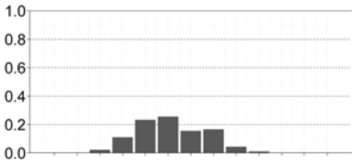
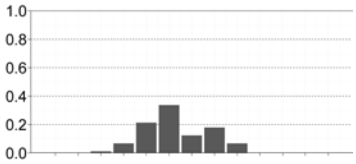
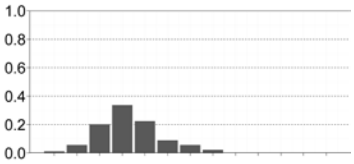
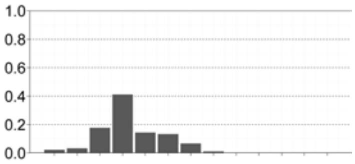
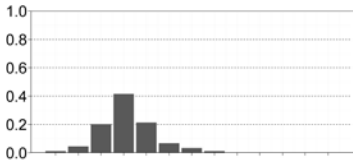
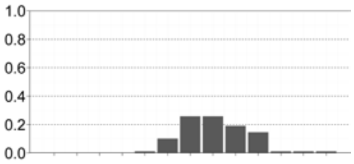
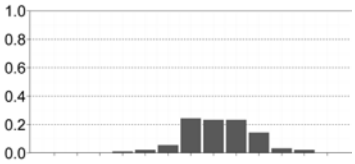
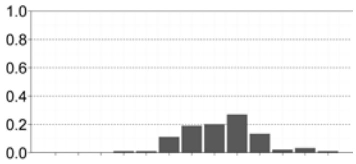
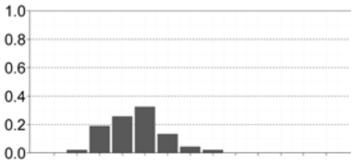
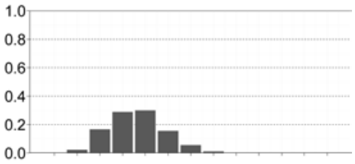
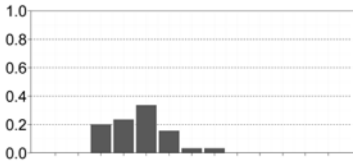
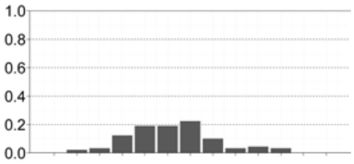
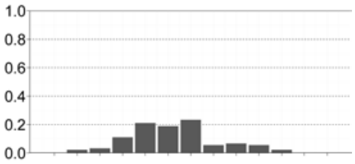
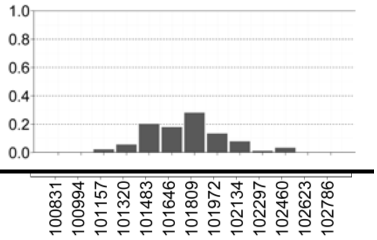
Distribution	Run 1	Run 2	Run 3
Data (training dataset)			
Model 1			
Model 2			
Model 3			
Model 4			

Table 9. Performance metrics (3) – (5) of probability distributions constructed from models 1 – 4 in the training period per cross validation run. Season OND, grid step size 162.89 Pa

Distribution	Run 1	Run 2	Run 3
		Mean	
Data (training dataset)	0% (abs. 101635 Pa)	0% (abs. 101669 Pa)	0% (abs. 101695 Pa)
Model 1	113%	-120%	-153%
Model 2	-153%	140%	144%
Model 3	94%	-102%	-116%
Model 4	-22%	7%	23%
		Standard deviation	
Data (training dataset)	0% (abs. 233.96 Pa)	0% (abs. 244.72 Pa)	0% (abs. 226.98 Pa)
Model 1	-5%	-10%	-14%
Model 2	1%	0%	17%
Model 3	-13%	-19%	-11%
Model 4	35%	29%	20%
		Statistical distance	
Data (training dataset)	0.000	0.000	0.000
Model 1	0.461	0.511	0.596
Model 2	0.584	0.533	0.494
Model 3	0.404	0.411	0.483
Model 4	0.180	0.200	0.225

4 Results

The prior probability distributions from models 1 – 4 are integrated using two alternative methods. We compute the product probability distribution using (1) and the BIC distribution using (2). The BIC weights in the linear combination of models 1 – 4 are dependent on the models' fit to the data in the training dataset. Table 10 presents the values of model weights in each cross validation run. The weights appear to be fairly stable across cross validation runs, but they vary strongly across seasons. Note that we do not use original time series to reconstruct probability mass functions of models 1 – 4 in the BIC method. Instead, the data points are applied to estimate parameters of the normally distributed priors (Renard et al., 2014), which are subsequently used in the BIC integration. However, we do not consider sensitivity of the results to the selected grid step size and compare the integrated BIC distribution with the discrete priors.

Table 10. BIC weights in cross validation runs.

	Model 1	Model 2	Model 3	Model 4
	Season JFM			
Run 1	0.6238	0.0019	0.0000	0.3743
Run 2	0.4101	0.0688	0.0000	0.5211
Run 3	0.5184	0.0917	0.0000	0.3899
	Season AMJ			
Run 1	0.0034	0.6942	0.0000	0.3024
Run 2	0.0000	0.6652	0.0000	0.3348
Run 3	0.0000	0.4728	0.0000	0.5272
	Season JAS			
Run 1	0.0036	0.9964	0.0000	0.0000
Run 2	0.1736	0.8264	0.0000	0.0000
Run 3	0.0001	0.9999	0.0000	0.0000
	Season OND			
Run 1	0.0000	0.0000	0.0000	1.0000
Run 2	0.0000	0.0000	0.0000	1.0000
Run 3	0.0000	0.0000	0.0000	1.0000

After that, we calculate performance metrics (3)-(5) separately for two probability distributions of instrumental observations. Firstly, we obtain the observed p_{data} from instrumental observations in the training dataset. Secondly, we calculate p_{data} from instrumental observations in the validation dataset. Tables 11-22 show histograms and performance estimates of the product and BIC distributions in each season and in each cross validation run.

Table 11. Integrated and data-based probability distributions per cross validation run. Season JFM, grid step size 229.87 Pa

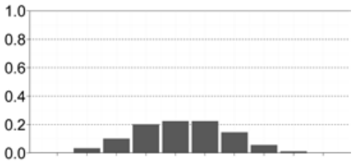
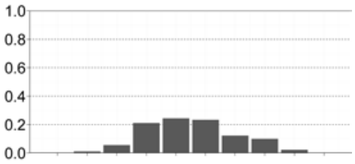
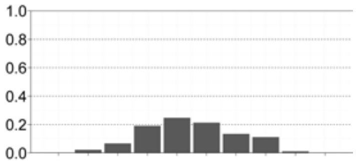
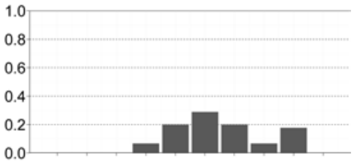
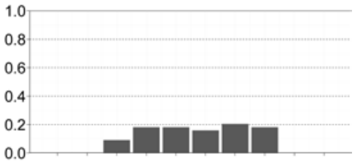
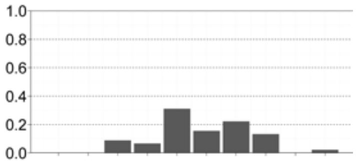
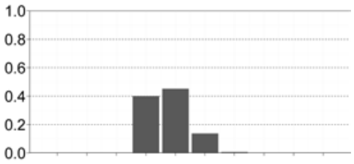
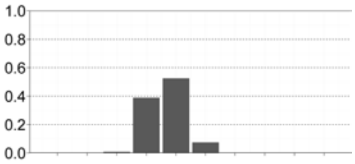
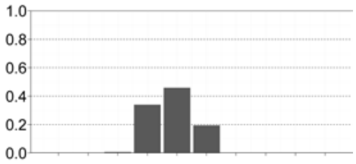
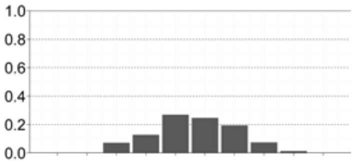
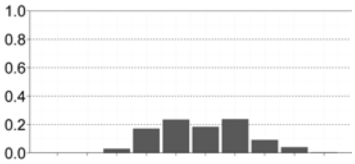
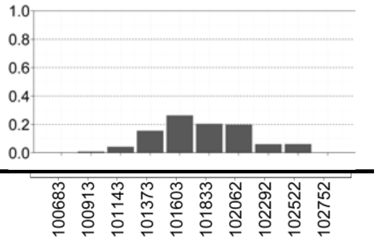
Distribution	Run 1	Run 2	Run 3
Data (training dataset)			
Data (validation dataset)			
Product			
BIC			

Table 12. Performance metrics (3) – (5) of the integrated probability distributions in the training period per cross validation run. Season JFM, grid step size 229.87 Pa

Distribution	Run 1	Run 2	Run 3
	Mean		
Data (training dataset)	0% (abs. 101654 Pa)	0% (abs. 101720 Pa)	0% (abs. 101711 Pa)
Product	-31%	-56%	-41%
BIC	27%	26%	20%
	Standard deviation		
Data (training dataset)	0% (abs. 352.68 Pa)	0% (abs. 344.80 Pa)	0% (abs. 354.93 Pa)
Product	-53%	-58%	-53%
BIC	-10%	0%	0%
	Statistical distance		
Data (training dataset)	0.000	0.000	0.000
Product	0.426	0.459	0.361
BIC	0.137	0.140	0.134

Table 13. Performance metrics (3) – (5) of the integrated probability distributions in the validation period per cross validation run. Season JFM, grid step size 229.87 Pa

Distribution	Run 1	Run 2	Run 3
	Mean		
Data (validation dataset)	0% (abs. 101955 Pa)	0% (abs. 101775 Pa)	0% (abs. 101802 Pa)
Product	-119%	-67%	-65%
BIC	-59%	9%	-6%
	Standard deviation		
Data (validation dataset)	0% (abs. 344.72 Pa)	0% (abs. 370.41 Pa)	0% (abs. 361.35 Pa)
Product	-52%	-61%	-53%
BIC	-7%	-7%	-2%
	Statistical distance		
Data (validation dataset)	0.000	0.000	0.000
Product	0.589	0.551	0.460
BIC	0.211	0.160	0.209

Table 14. Integrated and data-based probability distributions per cross validation run. Season AMJ, grid step size 103.4 Pa

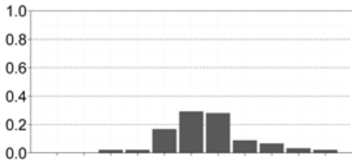
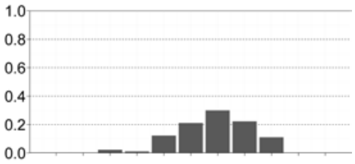
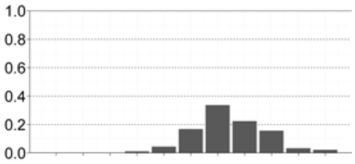
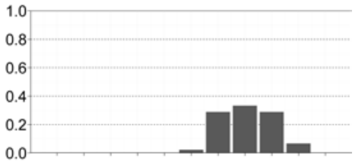
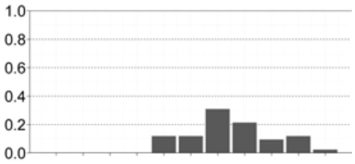
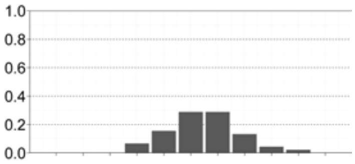
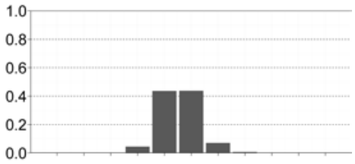
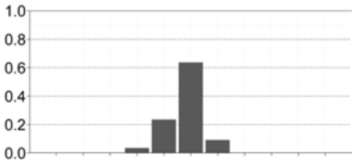
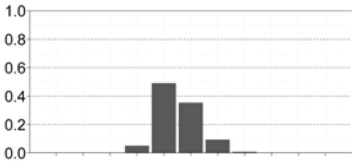
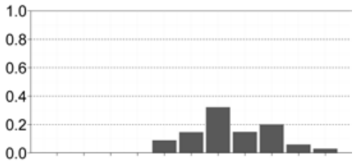
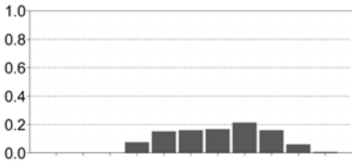
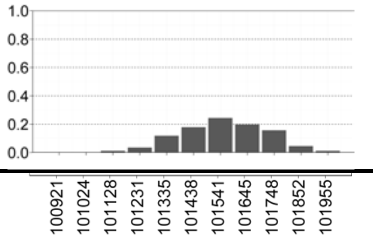
Distribution	Run 1	Run 2	Run 3
Data (training dataset)			
Data (validation dataset)			
Product			
BIC			

Table 15. Performance metrics (3) – (5) of the integrated probability distributions in the training period per cross validation run. Season AMJ, grid step size 103.4 Pa

Distribution	Run 1	Run 2	Run 3
		Mean	
Data (training dataset)	0% (abs. 101503 Pa)	0% (abs. 101528 Pa)	0% (abs. 101587 Pa)
Product	-69%	-80%	-144%
BIC	58%	14%	-23%
		Standard deviation	
Data (training dataset)	0% (abs. 160.63 Pa)	0% (abs. 138.90 Pa)	0% (abs. 137.73 Pa)
Product	-53%	-52%	-42%
BIC	-4%	29%	20%
		Statistical distance	
Data (training dataset)	0.000	0.000	0.000
Product	0.437	0.564	0.673
BIC	0.269	0.214	0.131

Table 16. Performance metrics (3) – (5) of the integrated probability distributions in the validation period per cross validation run. Season AMJ, grid step size 103.4 Pa

Distribution	Run 1	Run 2	Run 3
		Mean	
Data (validation dataset)	0% (abs. 101654 Pa)	0% (abs. 101593 Pa)	0% (abs. 101489 Pa)
Product	-264%	-111%	-73%
BIC	-58%	-29%	48%
		Standard deviation	
Data (validation dataset)	0% (abs. 99.47 Pa)	0% (abs. 159.95 Pa)	0% (abs. 137.22 Pa)
Product	-24%	-58%	-42%
BIC	55%	12%	21%
		Statistical distance	
Data (validation dataset)	0.000	0.000	0.000
Product	0.900	0.670	0.402
BIC	0.279	0.216	0.223

Table 17. Integrated and data-based probability distributions per cross validation run. Season JAS, grid step size 70.94 Pa

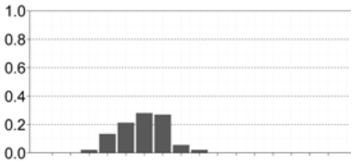
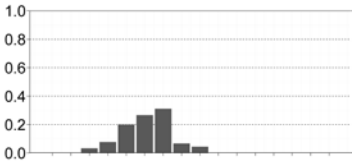
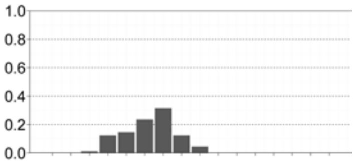
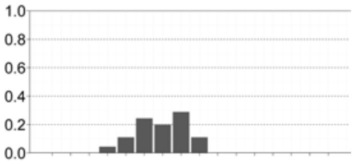
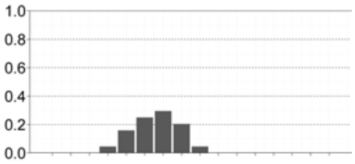
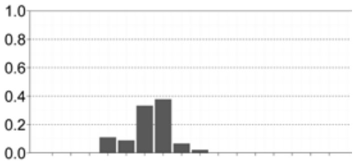
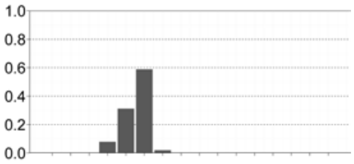
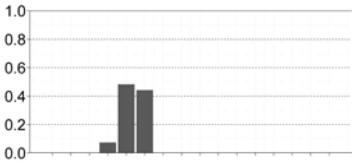
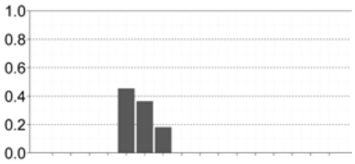
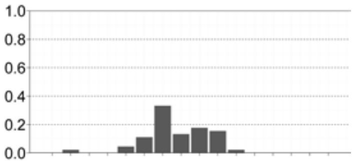
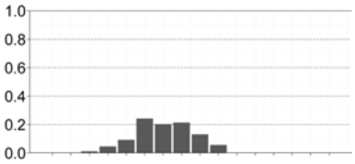
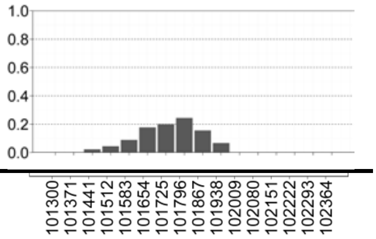
Distribution	Run 1	Run 2	Run 3
Data (training dataset)			
Data (validation dataset)			
Product			
BIC			

Table 18. Performance metrics (3) – (5) of the integrated probability distributions in the training period per cross validation run. Season JAS, grid step size 70.94 Pa

Distribution	Run 1	Run 2	Run 3
	Mean		
Data (training dataset)	0% (abs. 101647 Pa)	0% (abs. 101663 Pa)	0% (abs. 101673 Pa)
Product	-27%	-57%	-40%
BIC	143%	68%	66%
	Standard deviation		
Data (training dataset)	0% (abs. 90.89 Pa)	0% (abs. 94.48 Pa)	0% (abs. 97.33 Pa)
Product	-48%	-54%	-45%
BIC	33%	18%	20%
	Statistical distance		
Data (training dataset)	0.000	0.000	0.000
Product	0.407	0.459	0.437
BIC	0.494	0.291	0.309

Table 19. Performance metrics (3) – (5) of the integrated probability distributions in the validation period per cross validation run. Season JAS, grid step size 70.94 Pa

Distribution	Run 1	Run 2	Run 3
	Mean		
Data (validation dataset)	0% (abs. 101719 Pa)	0% (abs. 101696 Pa)	0% (abs. 101673 Pa)
Product	-101%	-99%	-47%
BIC	61%	36%	80%
	Standard deviation		
Data (validation dataset)	0% (abs. 95.55 Pa)	0% (abs. 87.30 Pa)	0% (abs. 81.09 Pa)
Product	-50%	-50%	-34%
BIC	27%	28%	44%
	Statistical distance		
Data (validation dataset)	0.000	0.000	0.000
Product	0.580	0.545	0.396
BIC	0.398	0.165	0.400

Table 20. Integrated and data-based probability distributions per cross validation run. Season OND, grid step size 162.89 Pa

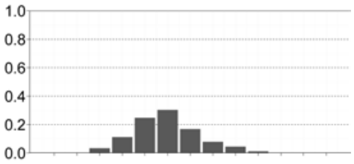
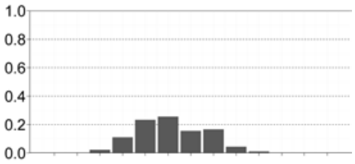
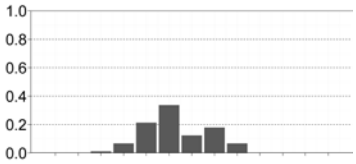
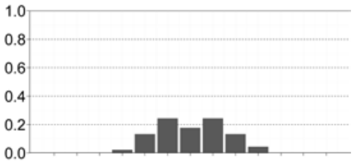
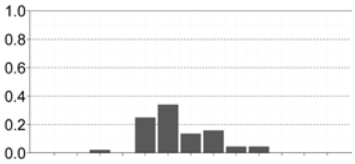
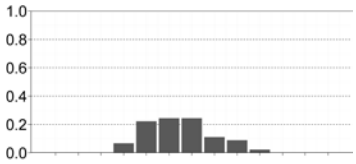
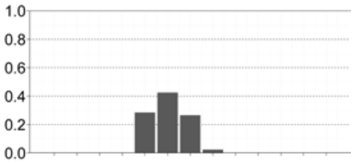
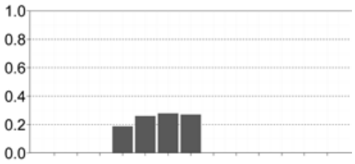
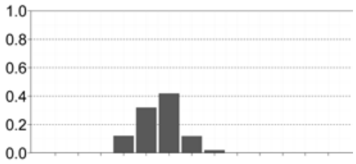
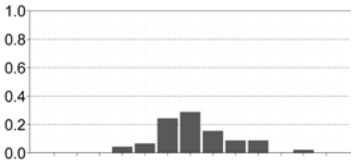
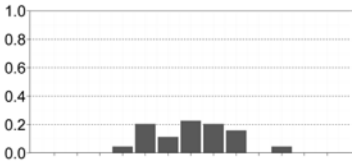
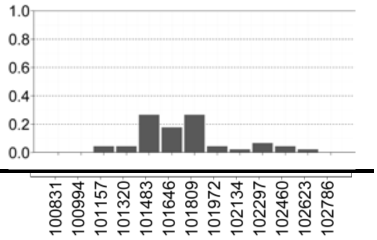
Distribution	Run 1	Run 2	Run 3
Data (training dataset)			
Data (validation dataset)			
Product			
BIC			

Table 21. Performance metrics (3) – (5) of the integrated probability distributions in the training period per cross validation run. Season OND, grid step size 162.89 Pa

Distribution	Run 1	Run 2	Run 3
	Mean		
Data (training dataset)	0% (abs. 101635 Pa)	0% (abs. 101669 Pa)	0% (abs. 101695 Pa)
Product	-7%	-34%	-51%
BIC	-88%	60%	18%
	Standard deviation		
Data (training dataset)	0% (abs. 233.96 Pa)	0% (abs. 244.72 Pa)	0% (abs. 226.98 Pa)
Product	-44%	-28%	-34%
BIC	15%	13%	46%
	Statistical distance		
Data (training dataset)	0.000	0.000	0.000
Product	0.257	0.242	0.242
BIC	0.341	0.270	0.363

Table 22. Performance metrics (3) – (5) of the integrated probability distributions in the validation period per cross validation run. Season OND, grid step size 162.89 Pa

Distribution	Run 1	Run 2	Run 3
	Mean		
Data (validation dataset)	0% (abs. 101820 Pa)	0% (abs. 101720 Pa)	0% (abs. 101722 Pa)
Product	-71%	-56%	-60%
BIC	9%	40%	6%
	Standard deviation		
Data (validation dataset)	0% (abs. 238.91 Pa)	0% (abs. 239.24 Pa)	0% (abs. 236.93 Pa)
Product	-45%	-27%	-37%
BIC	13%	16%	40%
	Statistical distance		
Data (validation dataset)	0.000	0.000	0.000
Product	0.420	0.333	0.327
BIC	0.200	0.341	0.222

At first, we assess whether integrated estimates reproduce the mean of the instrumental observations. In case of the posterior integration method, the average relative mean difference over seasonal cross validation runs drops only in the JAS season compared to the average model estimates before integration. On the other hand, the average relative mean difference over all runs, independent of season, is at the high rate of 53% for the training datasets and 94% for the validation datasets. In general, the BIC method can improve against the posterior integration method, if some of the models 1 – 4 are incorrect (e.g., the prior estimate may assign zero probability to the grid cell containing the true value; or the prior estimates are biased). Still the average difference in seasonal means of the BIC and observed distributions exceeds model ones (improving against the product mean in the JFM and AMJ seasons for the training dataset and in all seasons for the validation datasets). Additionally, the overall average relative mean difference is at the rate of 51% for the training datasets and 37% for the validation datasets.

Secondly, our goal is to compare standard deviations of the integrated distributions and the prior model distributions to assess whether knowledge about the true sea level pressure obtained from models 1 – 4 is improved after integration. The standard deviation criterion measures informativeness of the probability distributions; the less is the standard deviation of a probability distribution, the more informative it is (Kryazhimskiy et al., 2015). In this case study, the standard deviation of the integrated product distribution is less than the standard deviations of the sea level pressure distributions based on models 1 – 4 in each season and each cross validation run. Therefore, models 1 – 4 complement each other; the posterior integration method raises the informativeness of the model-based distributions. On the contrary, the standard deviation of the BIC distribution lies between the standard deviations of the model-based distributions in each season and in each cross validation run. In this case, we cannot argue about complementarity or disagreement of the model-based estimates (Kryazhimskiy et al., 2015).

However, the BIC method shows better fit to the observed distribution than the alternative approach of posterior integration. Here, we conducted analysis twice for the training and validation datasets, and measured fit in terms of the average relative difference in standard deviations between the integrated and observed distributions in all cross validation runs and in all cross validation runs per season. The averages were calculated from the raw seasonal estimates in Tables 12-13 for the JFM season, Tables 15-16 for the AMJ season, Tables 18-19 for the JAS season and Tables 21-22 for the OND season. The same conclusion holds true, when we change the comparative measure to the average statistical distance for both cases of training and validation datasets, except the case of training data in the OND season. A possible reason of these results is that the BIC method assesses the likelihood of instrumental observations to be generated from each of the model-based distributions, and subsequently, includes this information into the weights of the models' linear combination, which defines the integrated BIC distribution. Basically, BIC weights select models, which have better estimates of statistical distance than other models in the ensemble, but integration does not reduce this distance. On the contrary, the posterior integration method operates with the assumption that each prior distribution obtained from models 1 – 4 is equally likely to describe the unknown true sea level pressure; there is no ground to give a preference to the information obtained from any model in the ensemble.

In the end, we verify consistency of the integration results. For this purpose, we calculate the root mean squared error (6) between estimates measured relative to the training dataset and estimates measured relative to the validation dataset. For the relative difference in standard deviations between the product distribution and the observed distribution, the root mean squared error between results in the training and validation datasets equals 9% of the observed standard deviation in all cross validation runs. But this estimate varies in seasons, being around 2% for the runs in the JFM and OND seasons, 7% in the JAS season and 17% in the AMJ season. The consistency of the BIC results is twice worse than in the posterior integration case for each season and for all cross validation runs. At the same time, the root mean squared error for the statistical distance is 29% of the average seasonal distance in the training datasets in the JFM season, 57% in the AMJ season, 26% in the JAS season and 48% in the OND season in case of the product distribution. For the BIC distribution, the values equal 45%-26%-29%-38% respectively. These numbers with the estimates in the standard deviations in the JAS and AMJ seasons indicate significant difference in the results for the training and validation datasets for both integration methods.

Overall, after analysis of the integration results we anticipate that the errors in the probability distributions based on models 1 – 4 cannot be corrected by any of the integration methods studied in this paper.

Software

The estimates from the posterior integration method are obtained using the R package 'modelIntegration', which is available through the link http://www.iiasa.ac.at/web/home/research/researchPrograms/AdvancedSystemsAnalysis/modelIntegration-package.html at the International Institute for Applied Systems Analysis (IIASA).

References

[1] Burnham, K.P., Anderson, D.R. (2004). Multimodel inference - understanding AIC and BIC in model selection. Sociological Methods Research, 33(2): 261-304.

[2] Clemen, R.T. (1989). Combining forecasts: a review and annotated bibliography. International Journal of Forecasting, 5: 559-583.

[3] Compo, G.P. et al. (2011). The Twentieth Century Reanalysis Project. Quarterly Journal of the Royal Meteorological Society, 137(654): 1-28.

[4] Genest, C., Zidek, J. (1986). Combining probability distributions: a critique and an annotated bibliography. Statistical Science, 1(1): 114-135.

[5] Hastie T., Tibshirani R., Friedman J. (2001). The elements of statistical learning: data mining, inference and prediction. Springer-Verlag.

[6] Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T. (1999). Bayesian model averaging: a tutorial. Statistical Science, 14(4): 382-417.

[7] Kryazhimskiy, A.V. (2013). Posterior integration of independent stochastic estimates. IIASA Interim Report. IR-13-006.

[8] Kryazhimskiy, A. (2014). Systems analysis as a system of methods. How to study complex social-environmental systems? IIASA Interim Report. IR-14-021.

[9] Kryazhimskiy, A., Rovenskaya E., Shvidenko, A., Gusti, M., Shchepashchenko, D., Veshchinskaya, V. (2015). Towards harmonizing competing models: Russian foersts' net primary production case study. Technological Forecasting & Social Change, 98: 245-254.

[10] Kryazhimskiy, A.V. (2016). Posteriori integration of probabilities. Elementary theory. Theory of Probability and its Applications, 60(1): 62-87.

[11] Olson, D.L., Delen D. (2008). Advanced data mining techniques. Springer Publishing Company, Incorporated.

[12] Renard, B., Vidal, J.-P., Hingray, B., Raynaud, D. (2014). Quantifying uncertainty in climate projections using multimodel ensembles. COMPLEX report D2.5, 63p.

[13] Scott, D.W. (1979). On Optimal and Data-Based Histograms. Biometrika, 66(3): 605-610.

[14] Tebaldi, C., Knutti, R. (2007). The use of multi-model ensemble in probabilistic climate projections. Philosophical Transactions of the Royal Society A, 365: 2053-2075.