Web-based Supplementary Materials for "Monte Carlo inference for state-space models of wild animal populations" by Ken B. Newman, Carmen Fernández, Len Thomas, and Stephen T. Buckland

Web Appendix A: Application to simulated salmon data

We consider a Chinook salmon (*Oncorhynchus tshawytscha*) population in the Sacramento River of California. This population has suffered a drastic decline during the last few decades and was classified as endangered in 1994. One key management objective is to develop a means of combining outmigrating juvenile and adult return information in an effort to understand better the relationship between adult returns and juvenile production. Newman et al. (2006) fit SSMs to the available juvenile and adult data, and additional details of the problem background are included therein. Here, we present results of an analysis with simulated data based on the SSMs described in Newman et al. (2006). Simulated data were used instead of real data so as to provide a means of comparing the SIS and MCMC results to known values.

1 SSM formulation

State process: The state process is based on the following life history processes. Spawning occurs during September through December and fry emerge between January and March. Following a few months of freshwater residence, juvenile salmon smoltify and outmigrate to the Pacific ocean where they typically spend one to three years before returning to their natal stream to spawn and then die. Males tend to mature at a younger age than females. The time resolution of the model was annual with new time periods starting after spawning.

The distributions for the 11 components of the state vector are written below, where J_t are the number of outmigrating juveniles (age 1 fish), S_{ast} are the number of age a (a=2,3,4), sex s (s=male,female) spawning fish, and O_{ast} are the number of age a, sex s immature fish that remain in the ocean. With the exception of the juvenile salmon, Bernoulli processes are assumed for survival and maturation which then lead to binomial

and multinomial distributions.

$$J_t = (S_{\cdot f(t-1)})R_t, \quad \text{where } R_t \sim \text{Poisson}\left\{\frac{\alpha}{1+\beta S_{\cdot f(t-1)}}\right\} \tag{1}$$

$$(O_{2ft}, S_{2ft}, O_{2mt}, S_{2mt}) \sim \text{Multinomial}\{\ J_{t-1}\,;$$

$$0.5\phi_{2t}(1-\rho_{2f}), 0.5\phi_{2t}\rho_{2f}, 0.5\phi_{2t}(1-\rho_{2m}), 0.5\phi_{2t}\rho_{2m}\ \} \tag{2}$$

$$(O_{3ft}, S_{3ft}) \sim \text{Trinomial}\left\{\ O_{2f(t-1)}\,;\ \phi_3(1-\rho_{3f}), \phi_3\rho_{3f}\right\} \tag{3}$$

$$(O_{3mt}, S_{3mt}) \sim \text{Trinomial}\left\{\ O_{2m(t-1)}\,;\ \phi_3(1-\rho_{3m}), \phi_3\rho_{3m}\right\} \tag{4}$$

$$S_{4ft} \sim \text{Binomial}\left\{\ O_{3f(t-1)}\,;\ \phi_4\right\} \tag{5}$$

$$S_{4mt} \sim \text{Binomial}\left\{\ O_{3m(t-1)}\,;\ \phi_4\right\} \tag{6}$$

where $S_{\cdot f(t-1)} = S_{2f(t-1)}+S_{3f(t-1)}+S_{4f(t-1)}$. ϕ_a is the probability that a fish alive at age $a-1$ survives to age a. Survival to age 2, ϕ_{2t}, is time dependent. Maturation probabilities are age and sex dependent and denoted by ρ_{as} (with $\rho_{4s} = 1$). J_t is a product of the number of mature females and a Poisson variable, R_t, which represents number of juveniles produced per female. The expected value of R_t is a Beverton-Holt stock-recruitment function (Quinn and Deriso 1999), where α is the number of juveniles produced per female in the absence of density dependence, the strength of which is controlled by β. While multiple sub-processes were present, e.g. gender assignment, survival, and maturation, the pdfs for the combined processes were simply multinomials.

Observation process: For each year t, observations were estimates of returning adults S_{ast} and outmigrating juveniles J_t assuming instream sampling:

$$y_{Jt} \sim \text{N}\left\{J_t, \sigma^2_{Jt}\right\} \tag{7}$$

$$(y_{2ft}, y_{2mt}, y_{3ft}, y_{3mt}, y_{4ft}, y_{4mt})' \sim \text{MVN}\left\{(S_{2ft}, S_{2mt}, S_{3ft}, S_{3mt}, S_{4ft}, S_{4mt})', \Sigma_{St}\right\}, \tag{8}$$

with independence between the two Gaussian distributions and assuming σ^2_{Jt} and Σ_{St} known for all t. The assumed variances and covariances were loosely based upon values for the

binomial in the case of J_t and the multinomial for the case of the S_{ast}. For example, to estimate juvenile abundance an in-river trap with a known capture efficiency, say p, catches x juveniles and $y_{Jt} = x/p$; x is binomial(J,p) and the variance of y_{Jt} is $(1/p)^2 Jp(1-p)$ which is used for σ^2_{Jt}. For the adult data, a sample of returns is generally taken and the fish are categorized by age and sex, thus the numbers in each category can be modeled by a multinomial distribution. The "observed" values then are expansions of the sample proportions by the estimated total return.

Simulated data: States and observations were simulated for a 30 year period with $S_{\cdot ft}$ for the years t= -4, -3, -2, -1 fixed at 300, 400, 350, and 500, respectively to reflect a gradually increasing population. Values of α, β, ϕ_3, and ϕ_4 are shown in Web Table 1; values of ϕ_{2t} were sampled from a Uniform(0.001, 0.008) distribution. Maturation probabilities were fixed at $\rho_{2f} = 0.07$, $\rho_{2m} = 0.15$, $\rho_{3f} = 0.50$ and $\rho_{3m} = 0.60$ reflecting prior knowledge that males mature earlier than females. Given the fixed values of $S_{\cdot ft}$ (t=-4, ..., -1) and the parameters, the state process equations (1)-(6) were used to generate the series of state vectors $\mathbf{n}_0$, $\mathbf{n}_1$, ..., $\mathbf{n}_T$. Given the states, observations were simulated independently from (7) and (8).

Prior distributions: The prior distributions for α, β, ϕ_3, and ϕ_4 (Web Table 1) were independent, relatively wide and centered away from the true parameter values. The prior distribution for ϕ_{2t} was Beta (1.55,308) for all t, with parameters chosen such the expected survival probability was 0.5% with a standard deviation of 0.4%. We assumed the maturation probabilities were known; in the absence of information about ocean abundances (O_{ast}), maturation and survival probabilities are confounded.

Equations (1)-(6) jointly correspond to the state process pdf $g_t(\mathbf{n}_t|\mathbf{n}_{t-1})$. To define the initial state pdf, $g_0(\mathbf{n}_0|\eta)$, $S_{\cdot ft}$ for $t = -4, -3, -2, -1$ were taken as independent Uniform(200,1600) and equations (1)-(6) subsequently applied until t=0 was reached.

2 SIS implementation.

The LW algorithm was implemented in a similar manner to that for seals. Oversampling for year t=1 (using the auxiliary particle filter and kernel smoothing) was carried out using initial sample sizes of 60,000 particles of $\mathbf{n}_0$ and η, which were repeatedly generated and subsampled, until a target sample size of 1.5 million particles was accumulated.

3 MCMC implementation.

Blocking and analytical integration were used to improve the efficiency of the MCMC sampler. Strong posterior correlation exists between the juvenile production parameters, α and β, as well as between each ϕ_{2t} and numbers in the corresponding cohort O_{2ft}, S_{2ft}, O_{2mt}, S_{2mt}, $O_{3f(t+1)}$, $S_{3f(t+1)}$, $O_{3m(t+1)}$, $S_{3m(t+1)}$, $S_{4f(t+2)}$ and $S_{4m(t+2)}$. The fact that the observation vector does not include components for the numbers at ocean (O_{ast}) aggravates the problem. Consequently, α and β were sampled jointly, and the ocean abundances were integrated out of the state vector. Working by cohort instead of year we obtain

$$\begin{aligned}(S_{2ft}, S_{2mt}, S_{3f(t+1)}, S_{3m(t+1)}, S_{4f(t+2)}, S_{4m(t+2)}) \sim \text{Multinomial } \{\ J_{t-1}\,; \\ 0.5\phi_{2t}\rho_{2f}, 0.5\phi_{2t}\rho_{2m}\ ,\ 0.5\phi_{2t}(1-\rho_{2f})\phi_3\rho_{3f}\ ,\ 0.5\phi_{2t}(1-\rho_{2m})\phi_3\rho_{3m}\ , \\ 0.5\phi_{2t}(1-\rho_{2f})\phi_3(1-\rho_{3f})\phi_4\ ,\ 0.5\phi_{2t}(1-\rho_{2m})\phi_3(1-\rho_{3m})\phi_4\ \}\end{aligned} \tag{9}$$

It is enough to consider the state equation for the juveniles, given in (1), together with (9). Combining these with the observation equations and the prior distributions, the posterior distribution for the parameters and the number of juveniles and spawners is obtained. Posterior inference on the O_{ast} can be carried out by direct simulation after the MCMC sampler has been run. By integrating out the O_{ast}, the dimension of the posterior distribution went from $12 \times T + 37$ to $8 \times T + 15$, with $7 \times T$ observations to do the fitting.

To sample (α, β) we first used a random walk Gaussian proposal for β on the logit scale and then drew a candidate value for α from its full conditional posterior distribution (which has a known gamma form) conditioning on the candidate value for β; the candidate pair for (α, β) was accepted or rejected jointly. This was followed by a Gibbs update for α (conditioning on everything, including β). The remaining parameters and all states were sampled individually. The conditional posterior density for each of the survival parameters was log-concave, so we used Gibbs sampling via Adaptive Rejection Sampling (Gilks et al. 1996). Since states must take integer values, their conditional posterior distributions were mass functions. Often, the range of values a state could take was finite because it was constrained by the values of other states. We considered a continuous version of the target mass function and found, numerically, the mode of the function and the curvature at the mode. Then, we considered a t distribution (with a pre-specified number of degrees of freedom and truncated to the appropriate range) and chose the location and the scaling to match exactly the mode and curvature at the mode of the target distribution. Our proposal distribution was a discretized version of this t distribution; note that this was an independence sampler. Acceptance probabilities were always above 95%.

4 Results

The results presented for the SIS implementation were based on 1.5 million particles and took about 1.4 hours computing time, while the MCMC results were based on a run with a burn-in period of length 3 million and 7 million additional sweeps and took 13.6 hours. However, much shorter MCMC runs, e.g. 30,000+70,000 draws, led to similar results. There was substantial serial autocorrelation along the MCMC paths for most states and parameters (around 0.4 after 35 lags). Nevertheless, the algorithm converged without apparent problems: running the algorithm many times with widely scattered starting values gave

indistinguishable output.

Web Table 1 shows the posterior means for the parameters and the state in year 10 as well as relative errors using SIS and MCMC based on T=10 and T=30 years of data. The effect of time series length is relevant for fisheries management questions. The SIS and MCMC posterior means for the parameters and states were relatively close, as were density plots of the posterior distributions (based on T=30) of the parameters (not shown here). Posterior correlation between α and β was very strongly positive ($r_{\alpha,\beta} > 0.99$), consequently the marginal posterior distributions of α and β were quite wide and relative errors were large (Web Table 1). The fact that ϕ_{2t} was year-specific led to reasonably high uncertainty *a posteriori* and relative errors that were larger than those for ϕ_3 and ϕ_4, which were accurately estimated. Relative errors for ϕ_{2t} (not reported) were greater at the beginning and the end of the time series due to incomplete data regarding the fate of the relevant cohorts. Time series length had a sizeable effect on the relative error of α, β, ϕ_3, ϕ_4 and most of the year 10 states (Web Table 1), the relative errors for 10 years of data being roughly twice those for 30 years of data with some exceptions.

For the simulation results shown here, the posterior means, and the relative errors, for states and parameters were quite similar for both methods (Web Table 1). However, for smaller sample sizes, namely 100,000 particles (SIS) or draws (MCMC, of which 30,000 were burn-in), multiple runs demonstrated that the across-runs coefficient of variation of the relative error for parameters and states could be 5 to 10 times greater for SIS.

Web Table 1. Posterior means for salmon model parameters and states in year 10 with $T = 10$ and $T = 30$ years of data compared to true values. Prior distributions: $\alpha \sim$Gamma(4.68, 1068) (mean = 4998.24), $\beta \sim$Beta(0.99,997), $\phi_3 \sim$Beta(5.06, 5.06), $\phi_4 \sim$Beta(5.80, 3.87). Relative errors: $RE = 100 \times \left[E\left\{(\eta - \eta_{\text{true}})^2 | \mathbf{y}_{1:T}\right\}\right]^{1/2} / \eta_{\text{true}}$.

		$T = 10$ years				$T = 30$ years			
		SIS		MCMC		SIS		MCMC	
Param	Truth	$\hat{\eta}$	RE	$\hat{\eta}$	RE	$\hat{\eta}$	RE	$\hat{\eta}$	RE
α	4000	5619	47.8%	5631	48.7%	4722	25.2%	4713	24.8%
β	0.0015	0.0023	64.5%	0.0023	65.8%	0.0018	31.4%	0.0018	30.8%
ϕ_3	0.60	0.66	13.6%	0.66	13.0%	0.62	6.3%	0.62	6.1%
ϕ_4	0.70	0.81	16.8%	0.81	17.4%	0.73	6.9%	0.73	5.8%
		$T = 10$ years				$T = 30$ years			
		SIS		MCMC		SIS		MCMC	
State	Truth	$\hat{n}$	RE	$\hat{n}$	RE	$\hat{n}$	RE	$\hat{n}$	RE
J_{10}	2,196,684	2,107,814	4.9%	2,107,359	4.9%	2,137,027	3.8%	2,134,691	3.8%
$O_{2f(10)}$	1,170	1,300	43.7%	1,298	43.7%	1,110	21.0%	1,104	18.9%
$S_{2f(10)}$	83	97	48.0%	97	48.1%	83	23.9%	83	21.7%
$O_{2m(10)}$	1,053	1,188	44.8%	1,187	44.8%	1,013	20.6%	1,010	18.8%
$S_{2m(10)}$	199	209	40.2%	209	40.3%	178	22.5%	179	20.8%
$O_{3f(10)}$	1,561	1,679	13.6%	1,684	13.8%	1,602	8.9%	1,580	8.9%
$S_{3f(10)}$	1,515	1,673	15.4%	1,679	15.7%	1,592	9.9%	1,577	9.9%
$O_{3m(10)}$	1,116	1,227	15.3%	1,231	15.6%	1,166	10.0%	1,151	9.7%
$S_{3m(10)}$	1,716	1,843	13.2%	1,848	13.4%	1,758	9.1%	1,737	8.8%
$S_{4f(10)}$	1,307	1,214	12.5%	1,212	12.4%	1,127	17.2%	1,121	16.9%
$S_{4m(10)}$	904	888	11.1%	887	10.9%	830	13.2%	821	13.5%

Web Appendix B: Detailed results of analysis of simulated seal data

For each of the five simulated seal data sets, SIS and MCMC were carried out for five runs with equivalent computing time (for MCMC it was B=N=1.5 million and for SIS it was 20 million particles), and one additional longer run (for MCMC it was B=5 million and N=15 million, and for SIS it was 300 million particles). Posterior means for the parameters for each of the runs and each of the simulated seal data sets are plotted in Web Figures 1-5. Consistent, but for the most part relatively slight, differences between MCMC and SIS within a given dataset are apparent when comparing the longer run results. The signs of the differences tended to be identical across data sets as well, e.g., the posterior means based on the longer runs were consistently larger for MCMC than for SIS for all the parameters except ϕ_j (where the long run means from SIS were larger) and γ_{dist} (where there was no consistent pattern). Close examination of the plots indicates that, for the shorter runs, the MC variation of SIS tended to exceed the MC variation of MCMC, with the exception of α and ϕ_{ad} where the variation was relatively equal.

Web Figures 6-9 show the posterior means and standard deviations for pups and age 6+ females for the first simulated dataset (results are similar for the other data sets) for the shorter SIS and MCMC runs. The MC variation of posterior means of pups (Figure 6) for the SIS estimates tended to be slightly greater than that for MCMC estimates. However, for the posterior means of age 6+ females (Figure 8) the MC variation was somewhat greater for MCMC than for SIS. Similarly for posterior standard deviations, the MC variation of SIS was somewhat greater for than MCMC for pups (Figure 7) with the reverse true for age 6+ females (Figure 9).

Figure 1: Posterior means for parameters from simulated seal data set 1 for separate SIS and MCMC runs. The shorter runs for each method are denoted by points (triangles for SIS and circles for MCMC). The longer run values are horizontal lines (dashed lines for SIS and solid lines for MCMC).

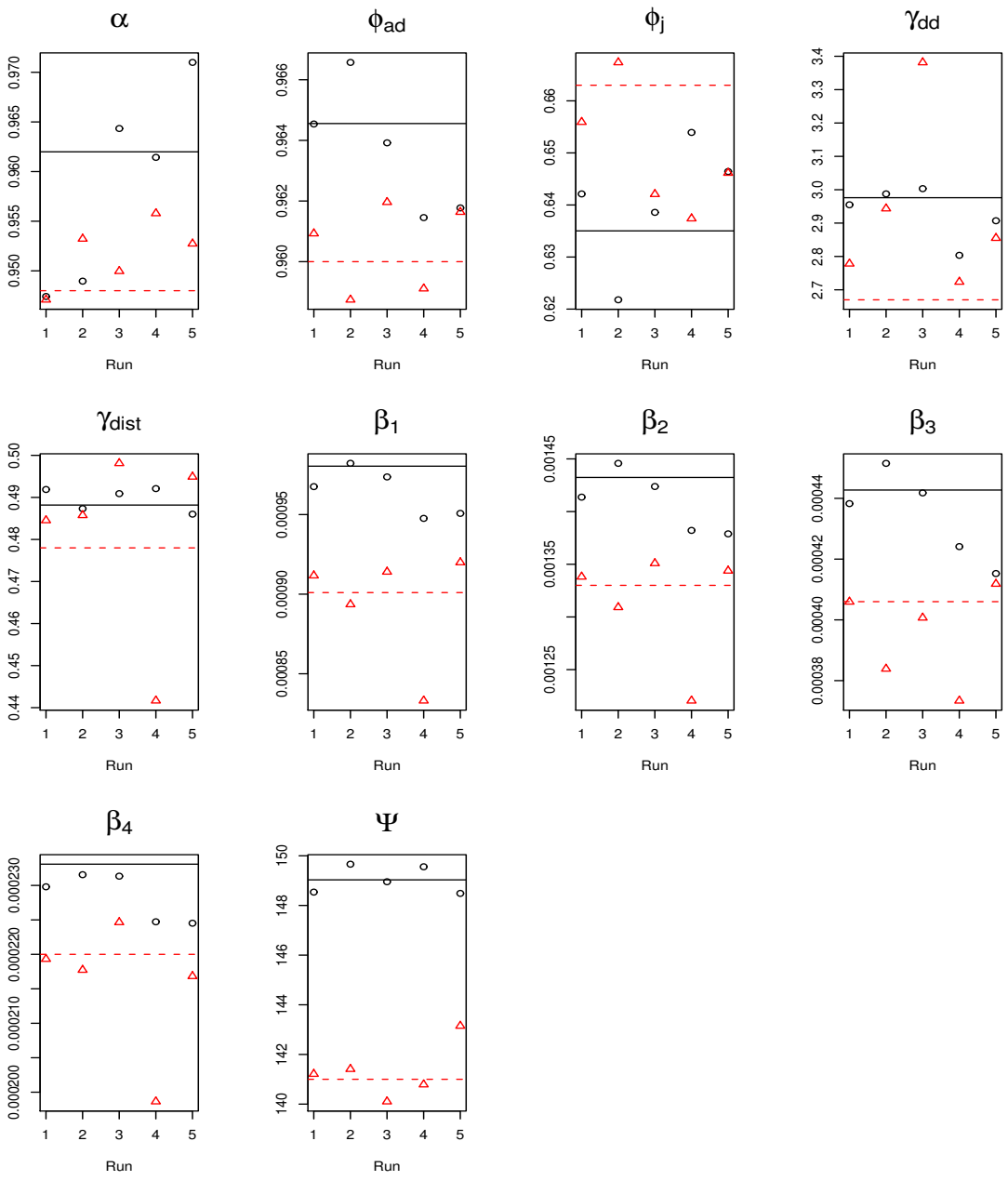

Figure 2: Posterior means for parameters from simulated seal data set 2 for separate SIS and MCMC runs. The shorter runs for each method are denoted by points (triangles for SIS and circles for MCMC). The longer run values are horizontal lines (dashed lines for SIS and solid lines for MCMC).

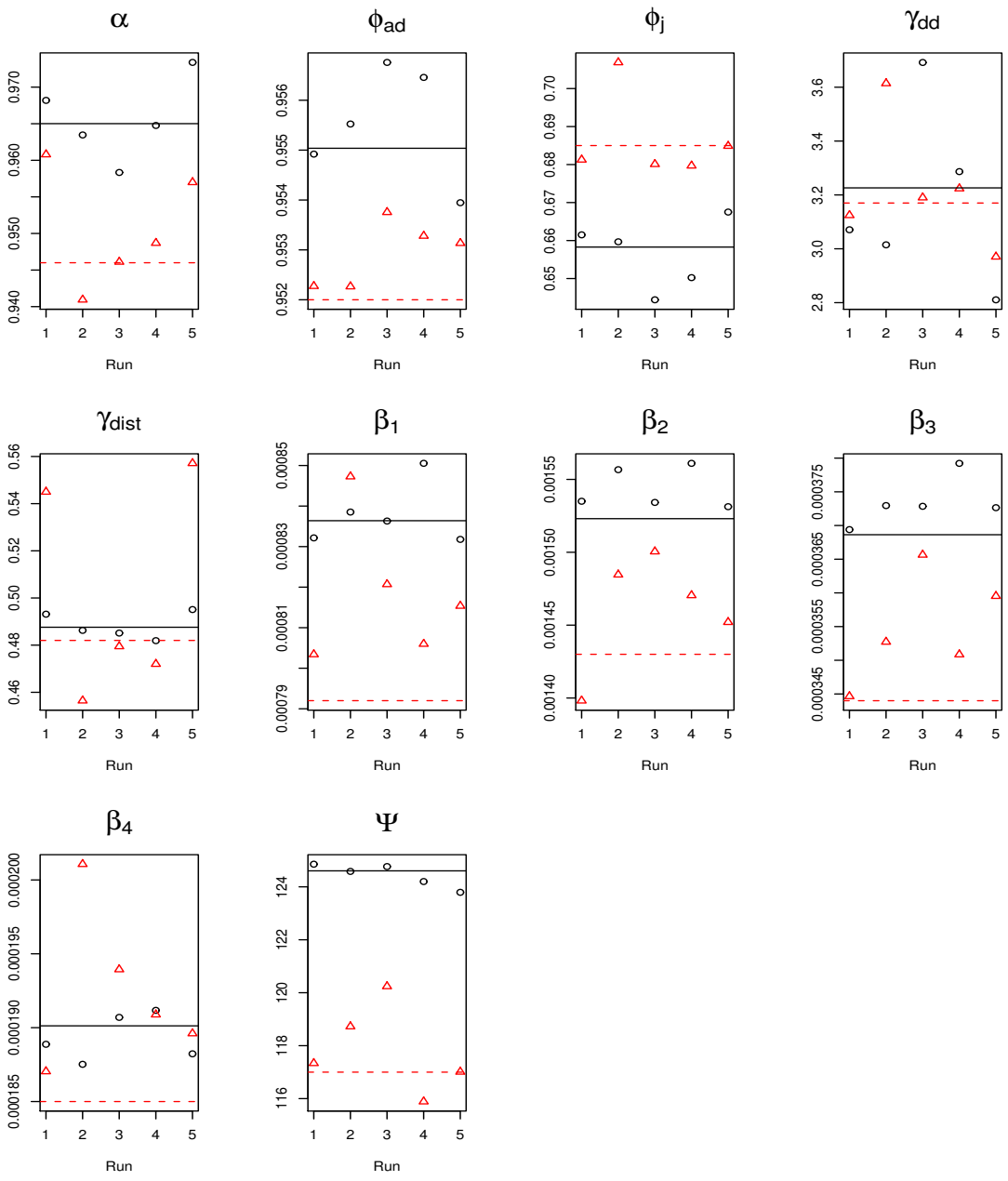

Figure 3: Posterior means for parameters from simulated seal data set 3 for separate SIS and MCMC runs. The shorter runs for each method are denoted by points (triangles for SIS and circles for MCMC). The longer run values are horizontal lines (dashed lines for SIS and solid lines for MCMC).

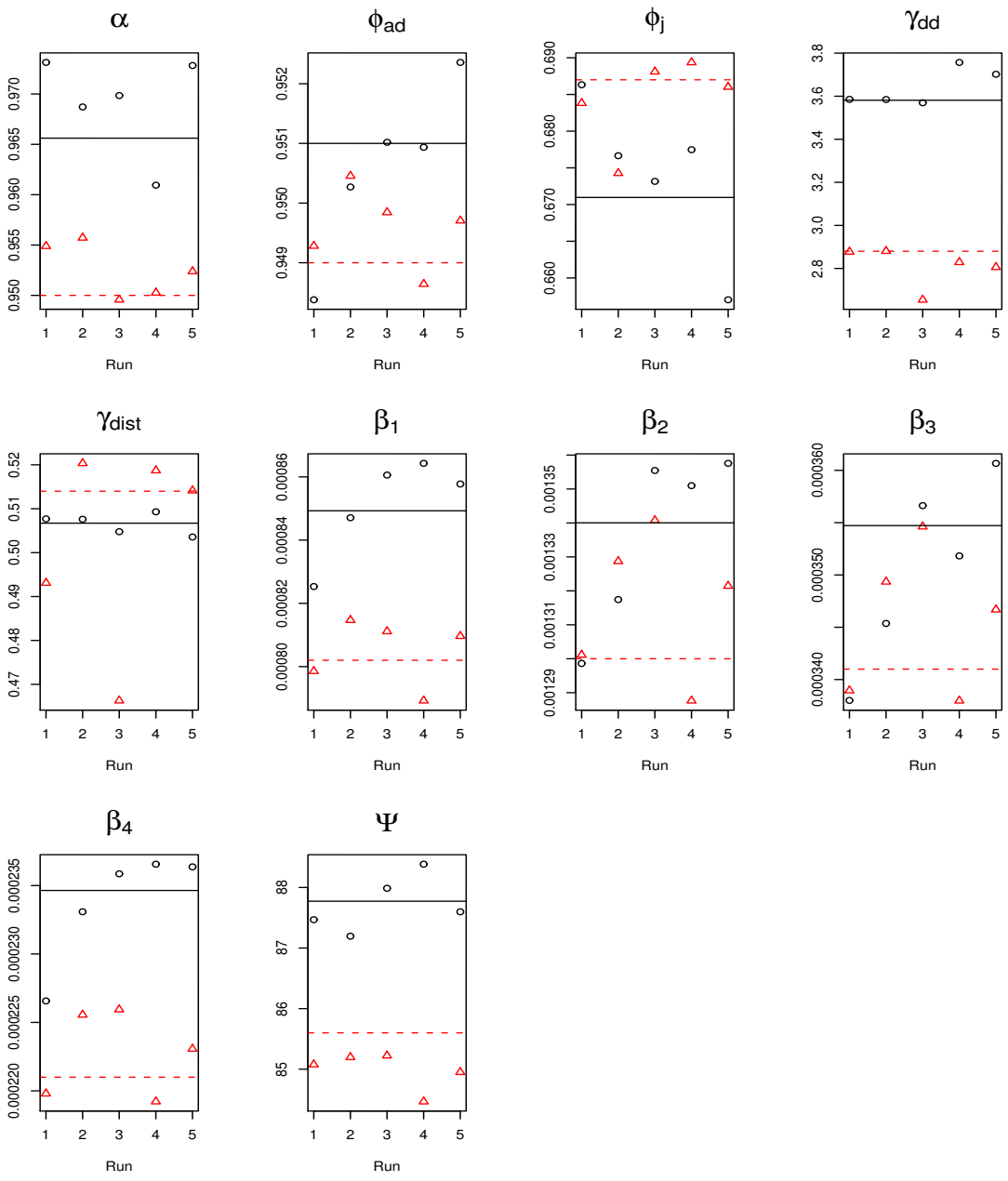

Figure 4: Posterior means for parameters from simulated seal data set 4 for separate SIS and MCMC runs. The shorter runs for each method are denoted by points (triangles for SIS and circles for MCMC). The longer run values are horizontal lines (dashed lines for SIS and solid lines for MCMC).

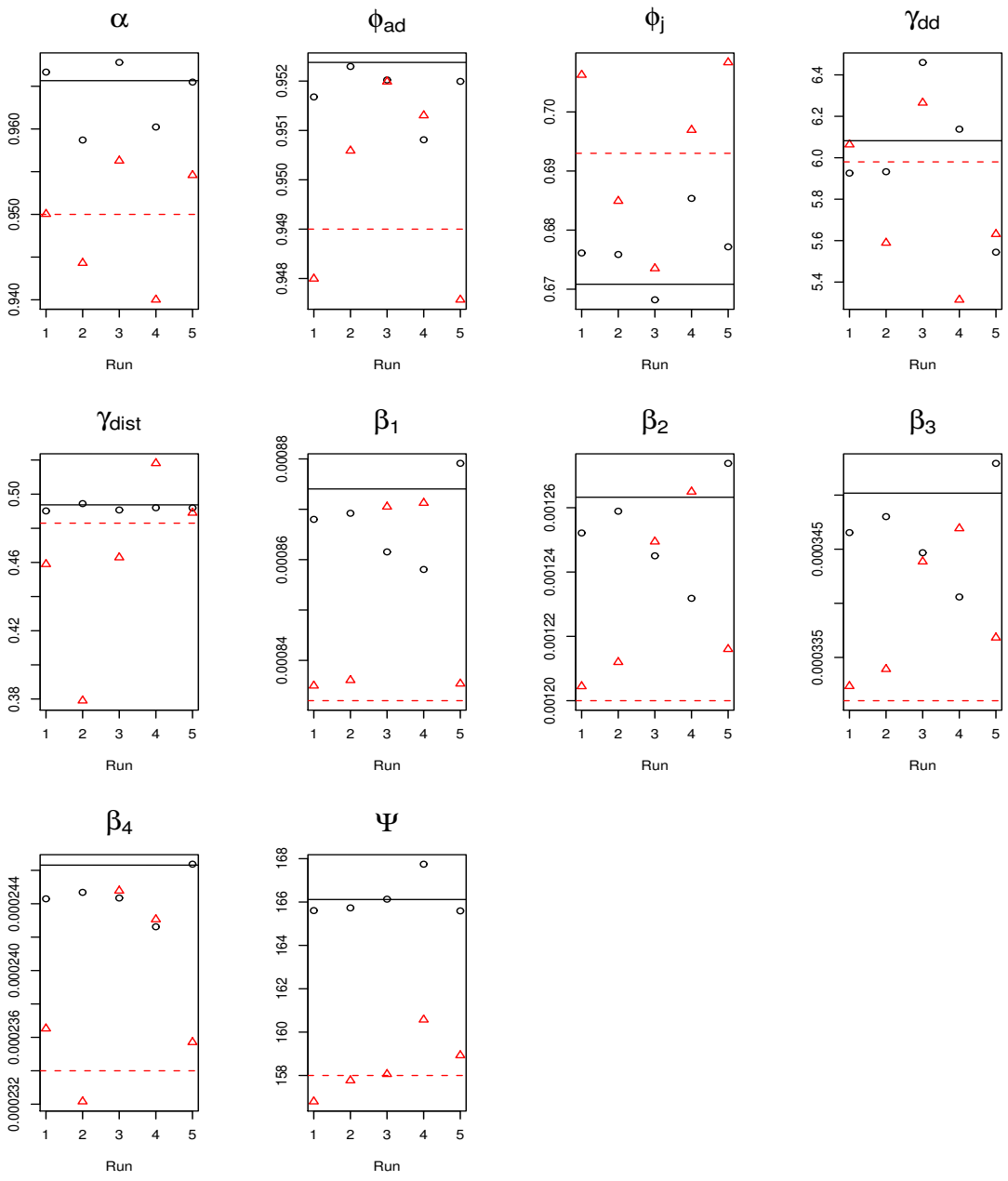

Figure 5: Posterior means for parameters from simulated seal data set 5 for separate SIS and MCMC runs. The shorter runs for each method are denoted by points (triangles for SIS and circles for MCMC). The longer run values are horizontal lines (dashed lines for SIS and solid lines for MCMC).

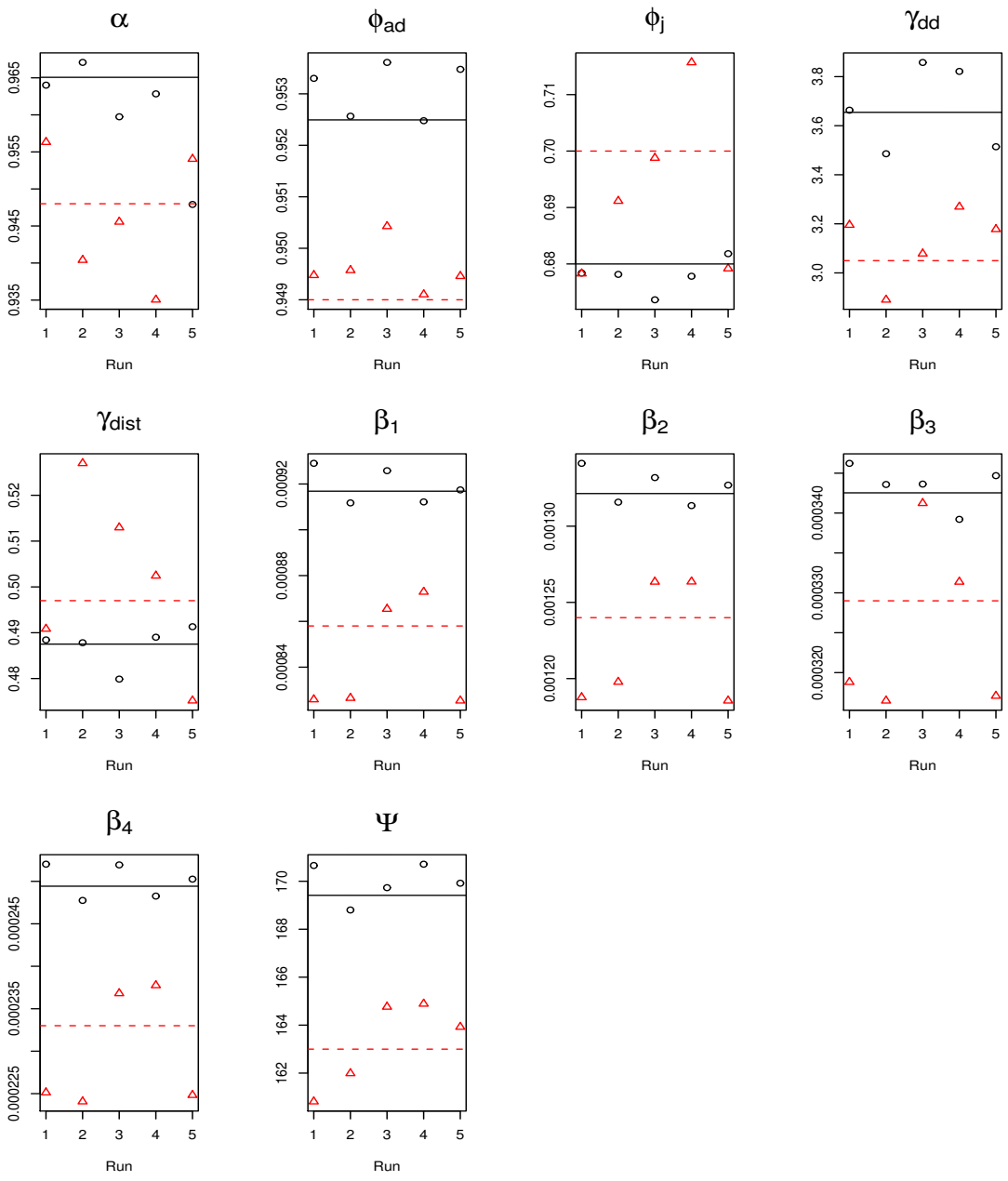

Figure 6: Posterior means for the number of pups from simulated seal data set 1 for separate SIS (triangles) and MCMC (circles) runs. Plot titles correspond to age (0), colony (1,2,3,4), and year (0,9,18), i.e., `nage.colony.year`. The longer run values are horizontal lines (dashed lines for SIS and solid lines for MCMC).

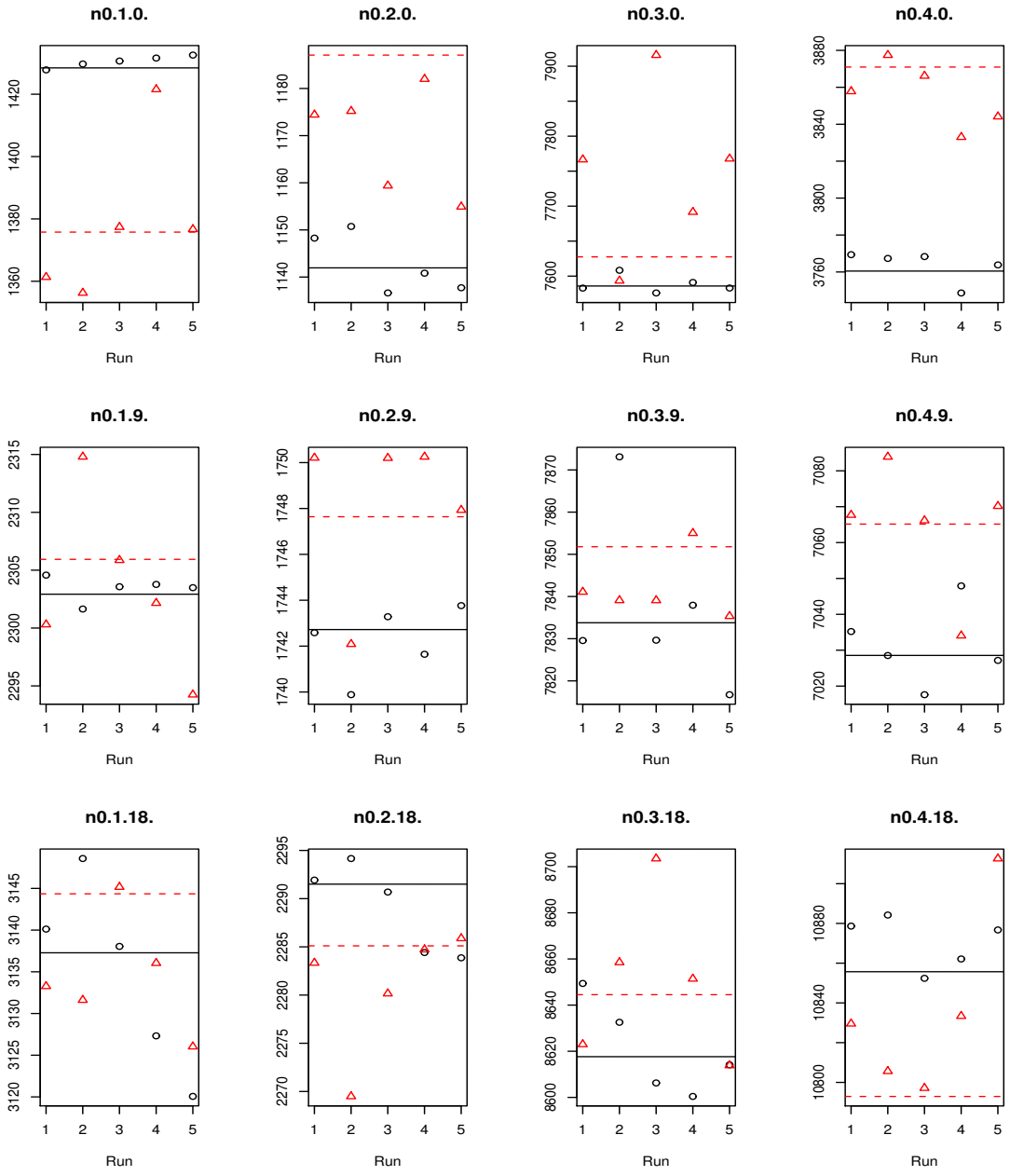

Figure 7: Posterior standard deviations for the number of pups from simulated seal data set 1 for separate SIS (triangles) and MCMC (circles) runs. Plot titles correspond to age (0), colony (1,2,3,4), and year (0,9,18), i.e., `nage.colony.year`. The longer run values are horizontal lines (dashed lines for SIS and solid lines for MCMC).

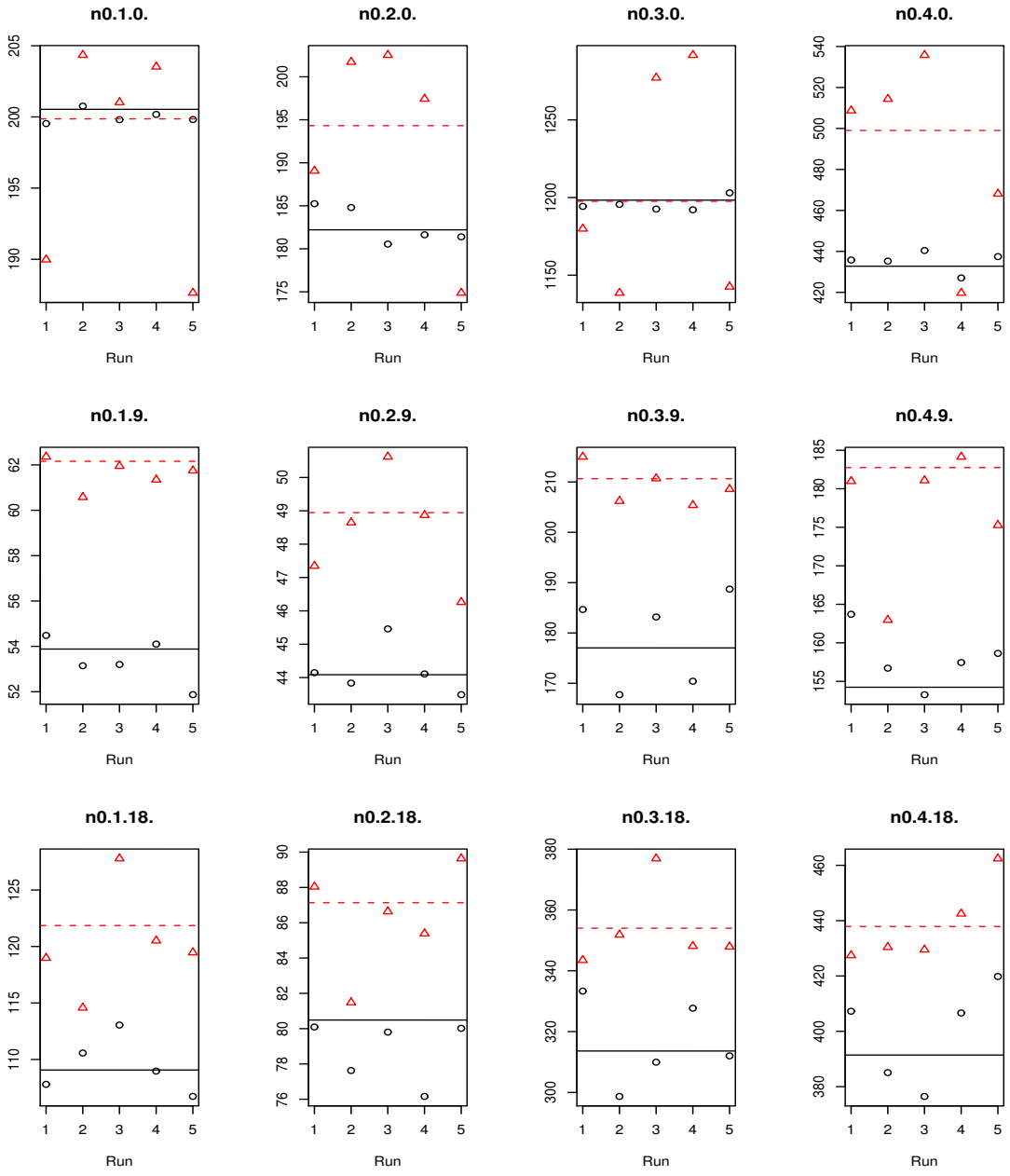

Figure 8: Posterior means for the number of age 6+ females from simulated seal data set 1 for separate SIS (triangles) and MCMC (circles) runs. Plot titles correspond to age (6+), colony (1,2,3,4), and year (0,9,18), i.e., `nage.colony.year`. The longer run values are horizontal lines (dashed lines for SIS and solid lines for MCMC).

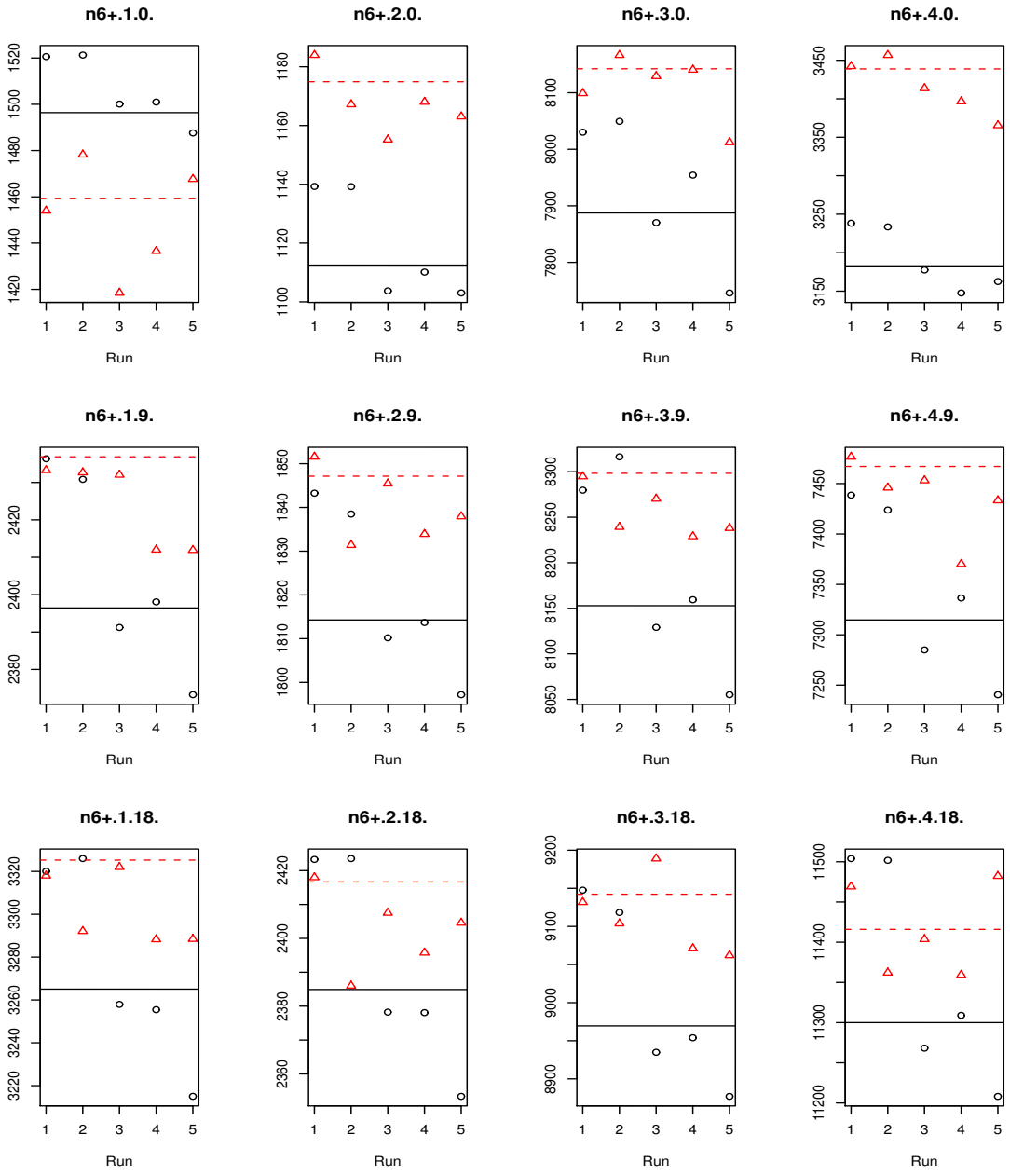

Figure 9: Posterior standard deviations for the number of age 6+ females from simulated seal data set 1 for separate SIS (triangles) and MCMC (circles) runs. Plot titles correspond to age (6+), colony (1,2,3,4), and year (0,9,18), i.e., `nage.colony.year`. The longer run values are horizontal lines (dashed lines for SIS and solid lines for MCMC).