Supporting Information for "A framework for variational inference and data assimilation of soil biogeochemical models using state space approximations and normalizing flows"

H. W. Xie[1,†], D. Sujono[2,†], T. Ryder[3], E. Sudderth[2], S. D. Allison[1,4]

[1]Center for Complex Biological Systems, University of California, Irvine, Irvine, CA, United States of America

[2]Department of Computer Science, University of California, Irvine, Irvine, CA, United States of America

[3]School of Mathematics Statistics and Physics, Newcastle University, Newcastle, United Kingdom

[4]Department of Ecology and Evolutionary Biology, University of California, Irvine, Irvine, CA, United States of America

[†]Authors contributed equally to this work.

Contents of this file

1. Figures S1 to S6
2. Table S1

Introduction

This document contains figures supporting the validity and functionality of our neural moving average flow VI framework. Figure S1 illustrates the benefit of initiating VI with an ELBO training warmup phase at low learning rates. Figure S2 demonstrates with an example $-\mathcal{L}$ trajectory from an SCON-C approximation inference that our VI algorithm is able to stably converge in ELBO. Figures S3 and S4 indicate that the neural moving

average flow VI approach remains viable for inference on approximated SCON-SS, and by extension, state space models that are linear in drift but non-linear in diffusion. Figure S5 depicts the importance of including CO_2 information in the data y for subtantial improvement of posterior identifiability and certainty. Figure S6 contrasts the effects of lengthening experiment time span T versus thickening observations in y to better inform and identify posteriors. Finally, Table S1 details the hyperparameters corresponding to our informed and independent univariate logit-normal priors.

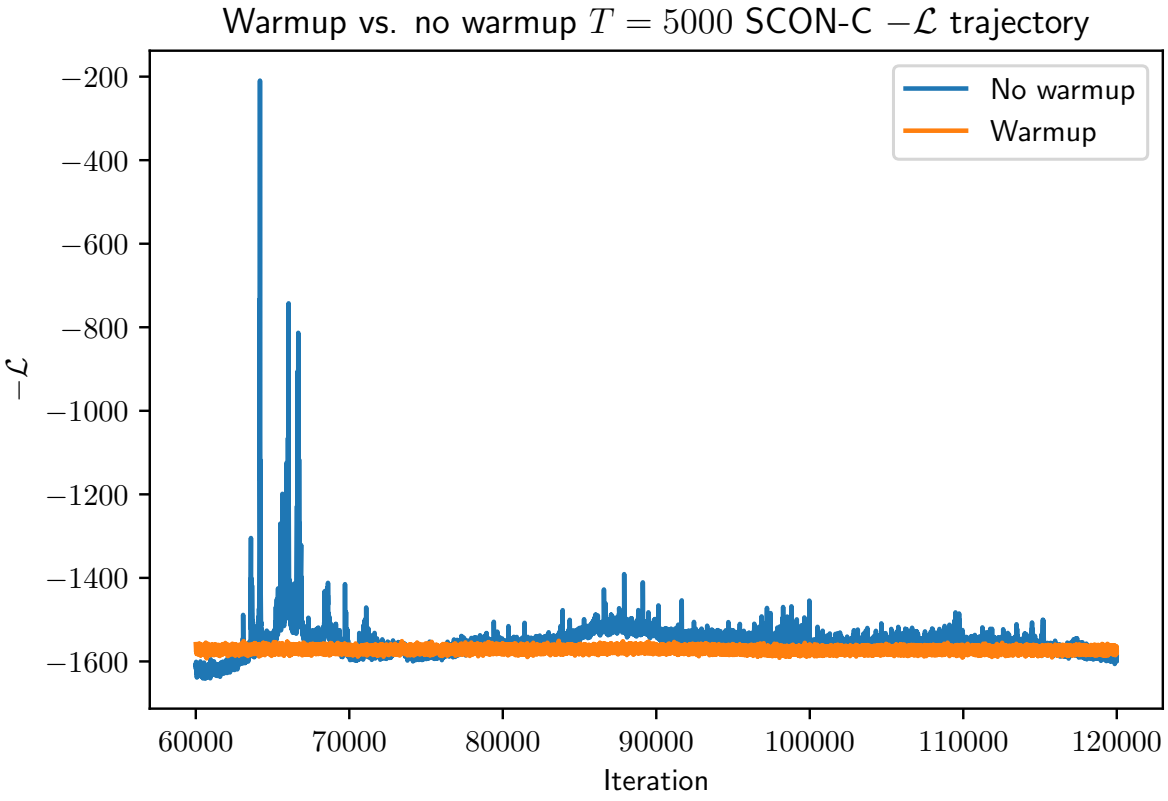

Figure S1. Comparison of $-\mathcal{L}$ trajectories from the latter halves of $T = 5000$ hour SCON-C flow trainings without (blue) and with (orange) warmup indicates that warmup helps stabilize training and speed up convergence. The trajectory corresponding to warmup displays much less prominent instability spiking and has flattened more quickly in contrast to the that of the no warmup counterpart.

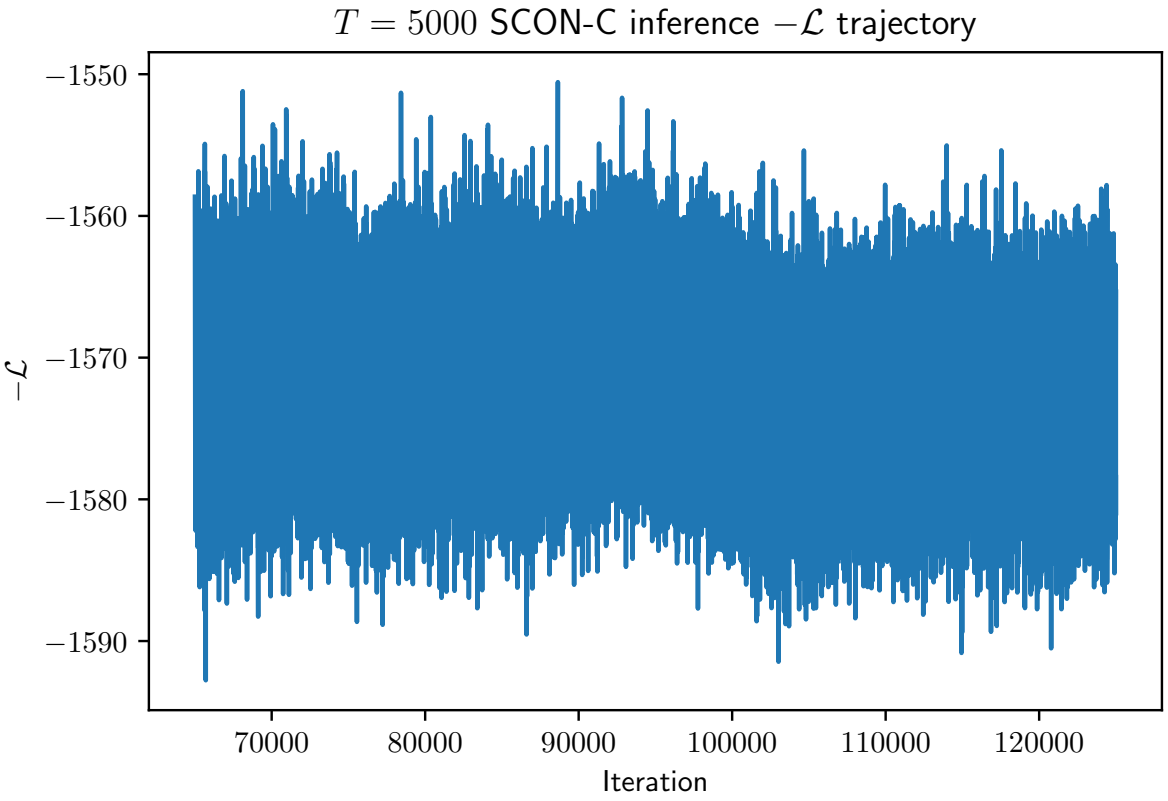

Figure S2. The stabilizing of the $-\mathcal{L}$ trajectory between -1550 and -1600 in the latter half of $T = 5000$ SCON-C flow VI training indicates convergence to an approximate local minimum $-\mathcal{L}$ and thereby proper algorithm function of the $q(\theta, x; \phi_{\theta,x})$ joint optimization.

Similarly stabilizing $-\mathcal{L}$ trajectories were observed for inferences on SCON-SS state space model approximations.

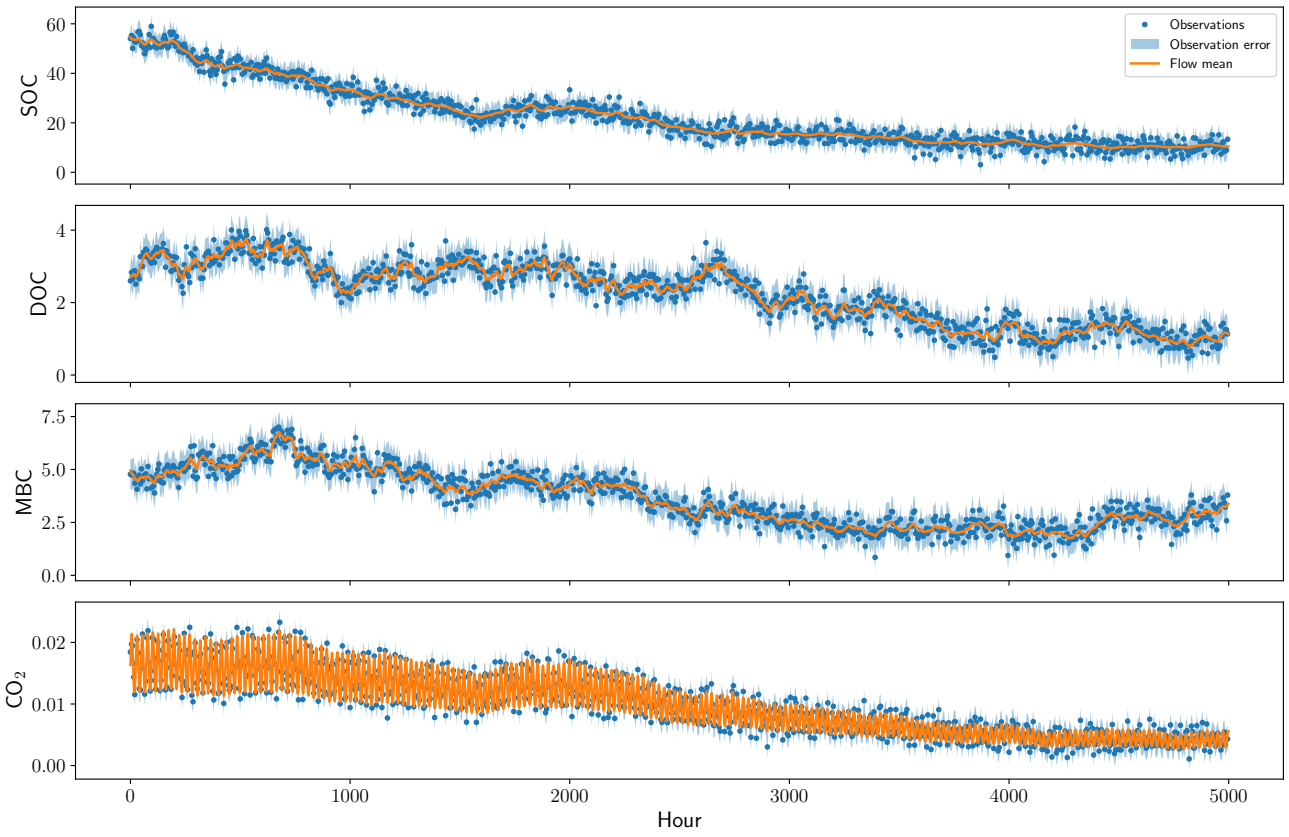

Figure S3. Flow-approximated SCON-SS $q(x|\theta;\phi_x)$ latent state and observed CO_2 means conditioned on $T = 5000$ SCON-SS data-generating process y estimated from 250 x paths sampled from the optimized joint variational $q(\theta, x; \phi_{(\theta,x)})$ density.

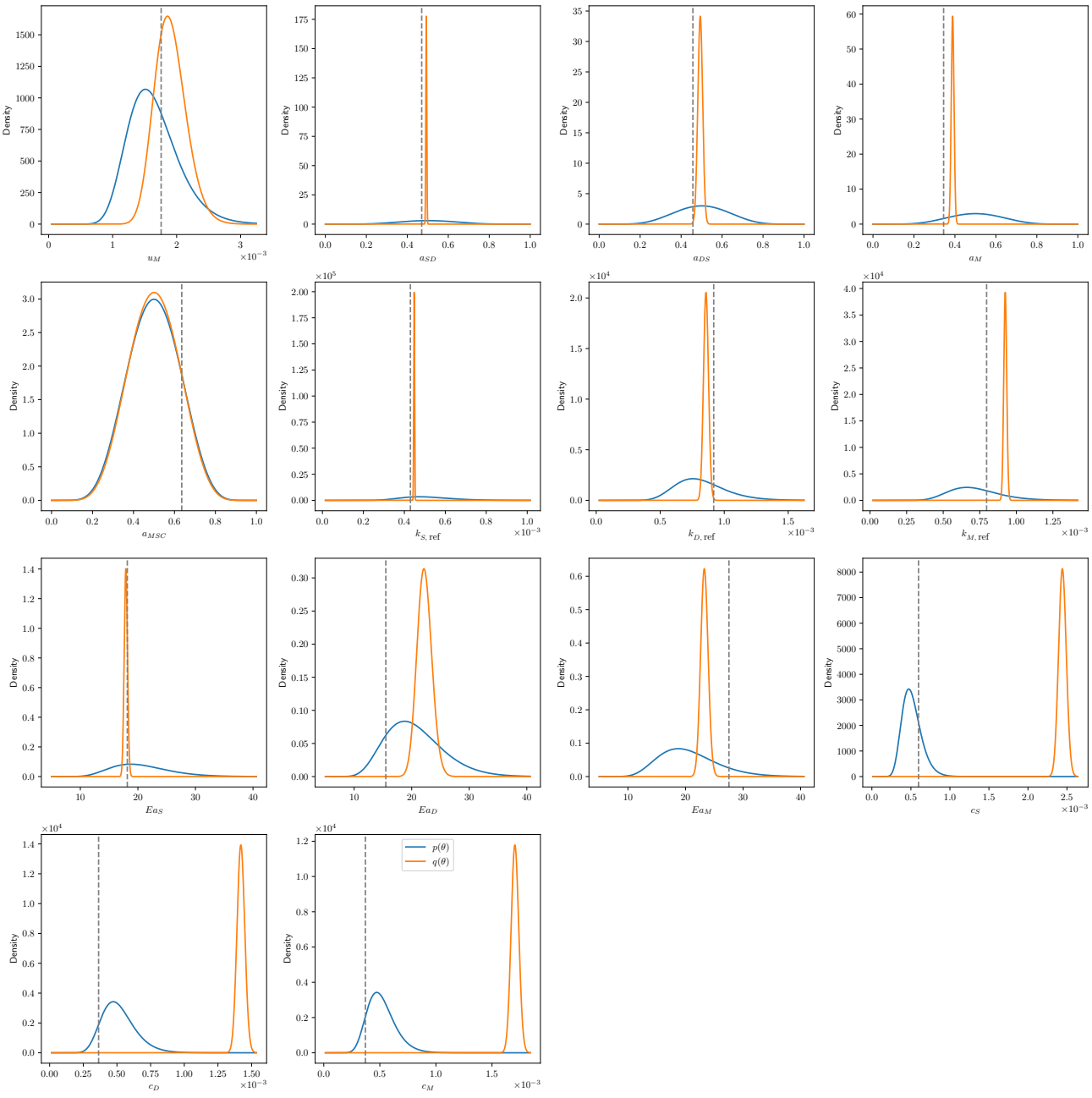

Figure S4. Full SCON-SS state space model marginal $q(\theta; \phi_\theta)$ posterior densities (orange) conditioned on $T = 5000$ SCON-SS data-generating process y compared to the prior densities $p(\theta)$ (blue). The true θ values sampled during data generation are marked by vertical dashed gray lines.

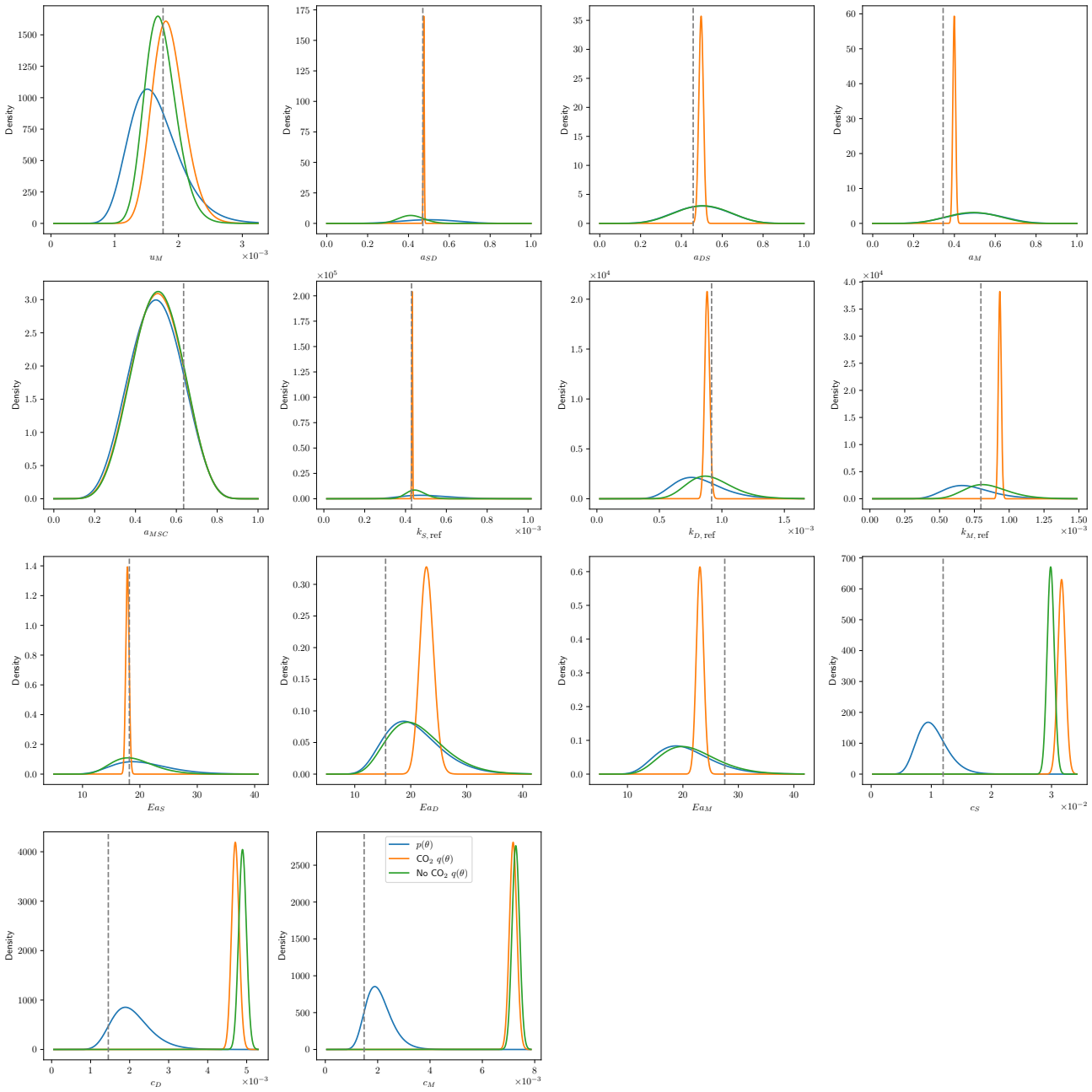

Figure S5. Approximate SCON-C state space model marginal $q(\theta;\phi_\theta)$ posterior densities conditioned with (orange) and without (green) CO_2 information in y produced by the same SCON-C data-generating process compared to mean-field prior densities $p(\theta)$ (blue). The true θ values sampled during data generation are marked by vertical dashed gray lines.

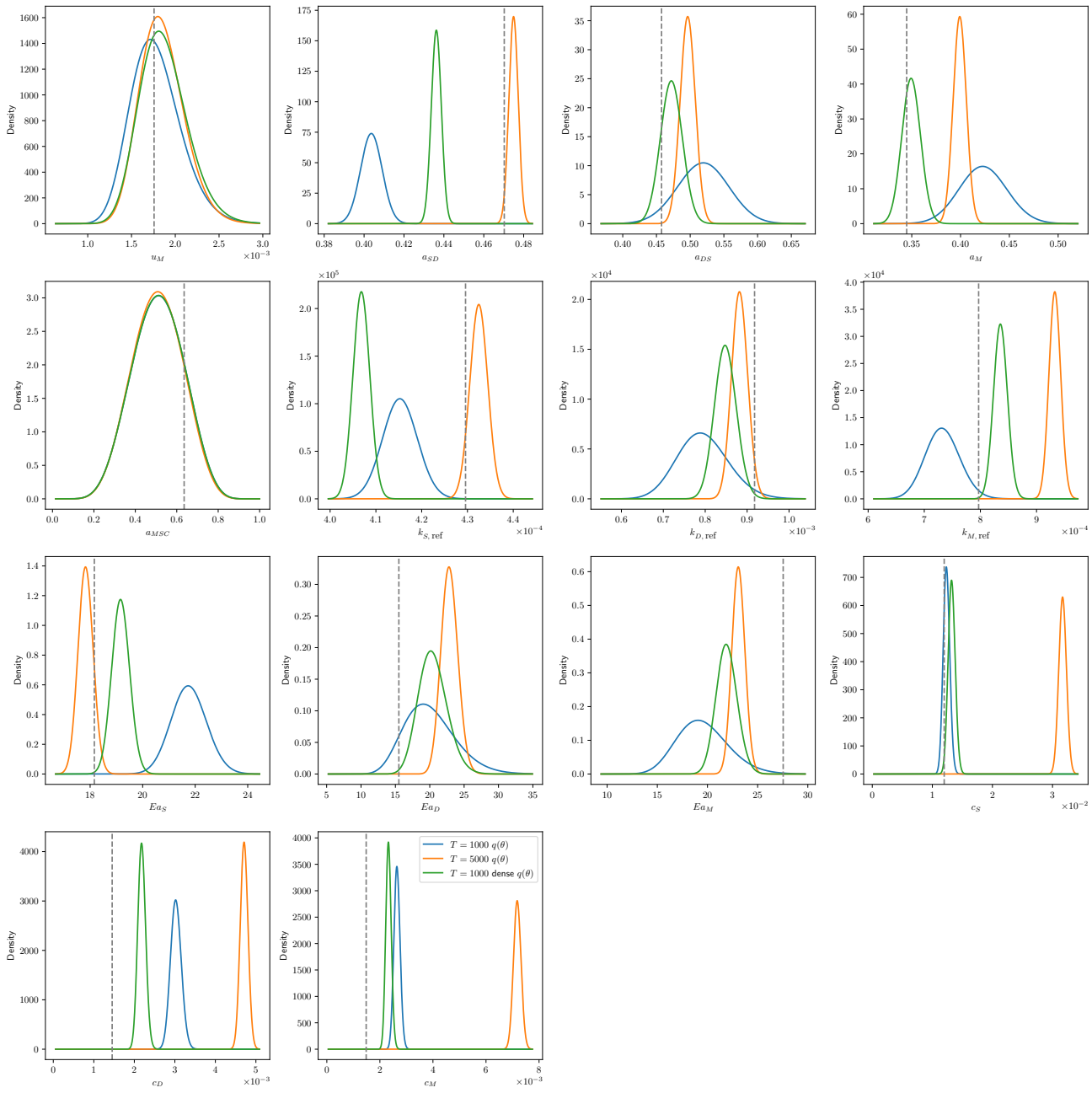

Figure S6. Approximate SCON-C state space model marginal $q(\theta; \phi_\theta)$ posterior densities conditioned with $T = 1000$ data observed every 5 hours (blue), $T = 5000$ data observed every 5 hours (orange), and $T = 1000$ data observed every hour (green). All three y share the same SCON-C data-generating process and include CO_2 information. The true θ values sampled during data generation are marked by vertical dashed gray lines.

θ	Biogeochemical interpretation	Target hyperparameters	Units
u_M	MBC uptake rate	$\mathscr{LN}(0.0016, 0.0004, 0, 1)$	$\mathrm{mg\,C\,g^{-1}\,C\,h^{-1}}$
a_{DS}	DOC to SOC transfer fraction	$\mathscr{LN}(0.5, 0.125, 0, 1)$	NA
a_{SD}	SOC to DOC transfer fraction	$\mathscr{LN}(0.5, 0.125, 0, 1)$	NA
a_M	MBC to organic C transfer fraction	$\mathscr{LN}(0.5, 0.125, 0, 1)$	NA
a_{MSC}	MBC to SOC transfer fraction	$\mathscr{LN}(0.5, 0.125, 0, 1)$	NA
$k_{S,\mathrm{ref}}$	SOC decomposition rate	$\mathscr{LN}(0.0005, 0.000125, 0, 0.1)$	$\mathrm{mg\,C\,mg^{-1}\,C\,h^{-1}}$
$k_{D,\mathrm{ref}}$	DOC decomposition rate	$\mathscr{LN}(0.0008, 0.0002, 0, 0.1)$	$\mathrm{mg\,C\,mg^{-1}\,C\,h^{-1}}$
$k_{M,\mathrm{ref}}$	MBC decomposition rate	$\mathscr{LN}(0.0007, 0.000175, 0, 0.1)$	$\mathrm{mg\,C\,mg^{-1}\,C\,h^{-1}}$
Ea_S	SOC decomposition activation energy	$\mathscr{LN}(20, 5, 5, 80)$	$\mathrm{kJ\,mol^{-1}}$
Ea_D	DOC decomposition activation energy	$\mathscr{LN}(20, 5, 5, 80)$	$\mathrm{kJ\,mol^{-1}}$
Ea_M	MBC decomposition activation energy	$\mathscr{LN}(20, 5, 5, 80)$	$\mathrm{kJ\,mol^{-1}}$
c_S	SCON-C SOC β constant	$\mathscr{LN}(0.1, 0.025, 0, 0.1)$	$\mathrm{mg\,C\,g^{-1}\,soil}$
c_D	SCON-C DOC β constant	$\mathscr{LN}(0.002, 0.0005, 0, 0.1)$	$\mathrm{mg\,C\,g^{-1}\,soil}$
c_M	SCON-C MBC β constant	$\mathscr{LN}(0.002, 0.0005, 0, 0.1)$	$\mathrm{mg\,C\,g^{-1}\,soil}$
s_S	SCON-SS SOC β factor	$\mathscr{LN}(0.0005, 0.000125, 0, 0.1)$	NA
s_D	SCON-SS DOC β factor	$\mathscr{LN}(0.0005, 0.000125, 0, 0.1)$	NA
s_M	SCON-SS MBC β factor	$\mathscr{LN}(0.0005, 0.000125, 0, 0.1)$	NA

Table S1. List of SCON-C and SCON-SS θ and their corresponding marginal data-generating and informed prior hyperparameters. The marginal densities are formatted as $\mathscr{LN}(\mu, \sigma, a, b)$, where μ and σ are the desired target density mean and standard deviation and a and b are the truncated distribution support lower and upper bounds.