

### Supplementary Information

Subsequent to the model tuning in this study, a large ensemble of climatology simulations (from October 1988 to September 2015) were run with PP set2 from the final selected 10 sets, with more than 100 simulations per year. Some initial analysis of the surface energy budget and surface radiative fluxes from this PP climatology were compared with a large ensemble of climatology simulations under SP to better understand the reduction in near surface temperature biases, shown in Fig. S16.

**Table S1.** Information of models used in Fig. 8, including the modelling institutions, model standard names, pertinent references, and ensemble members shown for each model.

Modeling institution	Model name	References	Ensemble member
Canadian Centre for Climate Modeling and Analysis	CanAM4	Chylek et al. (2011)	4
National Center for Atmospheric Research Community Earth System Model	CESM-CAM5	Neale et al. (2010)	2
Met Office Hadley Centre	HadGEM2-A	Martin et al. (2006) Collins et al. (2011)	6

**Figure S1.** Biases in a) June-July-August (JJA) mean temperature ( $^{\circ}\text{C}$ ), and b) precipitation (%) simulated by HadRM3P compared with PRISM over dec1996-nov 2007 under standard physics (SP) setting. The NWUS is defined as the land region bounded by the heavy grey line.

**Figure S2.** Phase 1 PPE parameter inputs and TOA outgoing SW and LW fluxes. 328 parameter sets are shown. The parameter values and model outputs under SP setting are marked in red.

**Figure S3.** Same as Fig. S3, but for Phase 2 parameter inputs and summary model output metrics considered in this phase. 264 parameter sets are shown.

**Figure S4.** Biases in a) June-July-August (JJA) mean temperature ( $^{\circ}\text{C}$ ) , and b) precipitation (%) simulated by HadRM3P compared with PRISM over dec1996-nov 2007 under the selected PP settings, where the composite of the final 10 are taken.

**Figure S5.** The range of internal variability for top-of-atmosphere a) outgoing shortwave radiation, b) outgoing longwave radiation, and c) net (outgoing minus incoming) under SP setting for each year. We rounded to the nearest  $\text{Wm}^{-2}$  ( $\pm 1$ ) to account for internal variability.

**Figure S6.** One-at-a-time sensitivity analysis of JJA temperature bias (compared with PRISM) over Northwest to each input parameter in turn, with all other parameters held at mean value of all the designed points. Central lines represent the emulator mean, and shaded areas represent the estimate of emulator uncertainty, at the  $\pm 1$  SD level.

**Figure S7.** Same as Fig. S6, but for DJF temperature bias.

**Figure S8.** Same as Fig. S6, but for JJA precipitation bias.

**Figure S9.** Same as Fig. S6, but for DJF precipitation bias.

**Figure S10.** Same as Fig. S6, but for TOA SW fluxes.

**Figure S11.** Same as Fig. S6, but for TOA LW fluxes.

**Figure S12.** Biases of SP temperature over land in a) DJF, b) MAM, c) JJA, and d) SON, compared with MERRA over December 1996 through November 2007. Biases of selected PP compared with MERRA are shown in e)-h), while the differences between selected PP and SP, i.e. the absolute increase or decrease of biases in PP with respect to the SP values, are shown in i) - l). The PP results are the composites of the 10 selected sets, 3 IC per set.

**Figure S13.** Same as Fig. S12, but for comparison with GHCN-CAMS.

**Figure S14.** Same as Fig. S12, but for comparison with CFSR.

**Figure S15.** Same as Fig. S12, but for comparison with NCEP.

**Figure S16.** MEAN summer (JJA) differences between SP and PPset2 for a) total downward shortwave radiation, and b) latent heat fluxes for the period Oct1988 – Sep2015.

**Figure S17.** Biases of SP precipitation over land in a) DJF, b) MAM, c) JJA, and d) SON, compared with GPCP over December 1996 through November 2007. Biases of selected PP compared with GPCP are shown in e)-h), while the differences between selected PP and SP, i.e. the absolute increase or decrease of biases in PP with respect to the SP values, are shown in i) - l). The PP results are the composites of the 10 selected sets, 3 IC per set.

**Figure S18.** Same as Fig. S17, but for comparison with GPCC.

**Figure S19.** Same as Fig. S17, but for comparison with MERRA.

**Figure S20.** Same as Fig. S17, but for comparison with ERAI.

**Figure S21.** Same as Fig. S17, but for comparison with JRA-55.

**Figure S22.** Same as Fig. S17, but for comparison with CFSR.

**Figure S23.** Same as Fig. S17, but for comparison with 20CRv2c.

## References:

- Chylek, P., Li, J., Dubey, M. K., Wang, M., and Lesins, G.: Observed and model simulated 20th century Arctic temperature variability: Canadian Earth System Model CanESM2, *Atmospheric Chemistry and Physics Discussions*, 11(8), 22,893–22,907. <https://doi.org/10.5194/acpd-11-22893-2011>, 2011.
- Collins, W. J., Bellouin, N., Doutriaux-Boucher, M., Gedney, N., Halloran, P., Hinton, T., Hughes, J., Jones, C.D., Joshi, M., Liddicoat, S. and Martin, G.: Development and evaluation of an Earth- System model— HadGEM2, *Geoscientific Model Development*, 4(4), 1051–1075. <https://doi.org/10.5194/gmd-4-1051-2011>, 2011.
- Martin, G. M., Ringer, M. A., Pope, V. D., Jones, A., Dearden, C., and Hinton, T. J.: The physical properties of the atmosphere in the new Hadley Centre Global Environmental Model (HadGEM1). Part I: Model description and global climatology, *Journal of Climate*, 19(7), 1274–1301. <https://doi.org/10.1175/JCLI3636.1>, 2006.
- Neale, R.B., Chen, C.C., Gettelman, A., Lauritzen, P.H., Park, S., Williamson, D.L., Conley, A.J., Garcia, R., Kinnison, D., Lamarque, J.F. and Marsh, D.: . Description of the NCAR Community Atmosphere Model (CAM 5.0). Tech. Rep. NCAR/TN-486+ STR, 1(1), pp.1-12., 2010.