Title
The Urban-PLUMBER model evaluation project: Phase 1 results

Authors
Mathew Lipson, Sue Grimmond, Martin Best, Gab Abramowitz, Andrew Pitman, Andrew Coutts, Nigel Tapper, Jong-Jin Baik, Meiring Beyers, Lewis Blunn, Souhail Boussetta, Elie Bou-Zeid, Martin G. De Kauwe, Cécile de Munck, Matthias Demuzere, Simone Fatichi, Krzysztof Fortuniak, Beom-Soon Han, Maggie Hendry, Yukihiro Kikegawa, Hiroaki Kondo, Doo-Il Lee, Sang-Hyun Lee, Aude Lemonsu, Tiago Machado, Gabriele Manoli, Alberto Martilli, Valéry Masson, Joe McNorton, Naika Meili, David Meyer, Kerry A. Nice, Keith W. Oleson, Seung-Bu Park, Michael Roth, Robert Schoetter, Andres Simon, Gert-Jan Steeneveld, Ting Sun, Yuya Takane, Marcus Thatcher, Aristofanis Tsiringakis, Mikhail Varentsov, Chenghao Wang, Zhi-Hua Wang

Abstract
Multi-model evaluations can help improve the skill of models and modellers, however no urban-focused land surface model intercomparison has be run since PILPS-Urban in 2011. Here, in Phase 1 of the Urban-PLUMBER project, we evalute 30 land surface models on their ability to simulate surface energy fluxes – critical inputs to atmospheric meteorological and air quality simulations. We use benchmarks (simple, information limited models) to set minimum and maximum performance expectations. We compare results directly with PILPS-Urban, undertaken at the same site (Melbourne, Australia). Overall, we find broad improvement in the current cohort's predictions of shortwave radiation, sensible and latent heat fluxes, but little or no improvement in longwave radiation and momentum fluxes. We find that efforts to integrate vegetation and hydrological processes into mid-complexity canyon models has paid dividends, with many of them performing as well as the simpler one and two-tile urban schemes. In contrast with recent non-urban intercomparisons, we find participating models generally perform well compared with the simple empirically based benchmarks. We also identified widespread human errors in initial submissions that substantially affected model performances. Although significant efforts were applied to correct these errors, we conclude that human factors are likely to influence results in this (or any) model intercomparison, particularly where the experience of participating scientists varies. We make the observational forcing, analysis, and benchmark data openly available, and use a web-based model evaluation portal, providing the community tools to test and improve the skill of future models and future modellers.