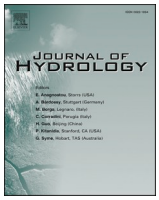
Journal of Hydrology 661 (2025) 133631

Contents lists available at ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol



# Evaluation of multimodel averaging approaches for ensembling evapotranspiration and yield simulations from maize models

Viveka Nand [a], Zhiming Qi [a,*], Liwang Ma [b], Matthew J. Helmers [c], Chandra A. Madramootoo [a], Ward N. Smith [d], Tiequan Zhang [e], Tobias K.D. Weber [f], Elizabeth Pattey [d], Ziwei Li [a], Jiaxin Wang [a], Virginia L. Jin [g], Qianjing Jiang [h], Mario Tenuta [i], Thomas J. Trout [j], Haomiao Cheng [k], R. Daren Harmel [l], Bruce A. Kimball [m], Kelly R. Thorp [m], Kenneth J. Boote [n], Claudio Stockle [o], Andrew E. Suyker [p], Steven R. Evett [q], David K. Brauer [q], Gwen G. Coyle [q], Karen S. Copeland [q], Gary W. Marek [q], Paul D. Colaizzi [q], Marco Acutis [r], Seyyed Majid Alimagham [s], Sotirios Archontoulis [t], Faye Babacar [u], Zoltán Barcza [v,w], Bruno Basso [x], Patrick Bertuzzi [y], Julie Constantin [z], Massimiliano De Antoni Migliorati [aa], Benjamin Dumont [ab], Jean-Louis Durand [ac], Nándor Fodor [ad], Thomas Gaiser [ae], Pasquale Garofalo [af], Sebastian Gayler [ag], Luisa Giglio [af], Robert Grant [ah], Kaiyu Guan [ai], Gerrit Hoogenboom [n], Soo-Hyung Kim [aj], Isaya Kisekka [ak], Jon Lizaso [al], Sara Masia [am], Huimin Meng [an], Valentina Mereu [ao], Ahmed Mukhtar [ap,aq], Alessia Perego [r], Bin Peng [ai], Eckart Priesack [ar], Vakhtang Shelia [n], Richard Snyder [as], Afshin Soltani [g], Donatella Spano [as], Amit Srivastava [ae], Aimee Thomson [ah], Dennis Timlin [au], Antonio Trabucco [ao], Heidi Webber [av], Magali Willaume [z], Karina Williams [aw], Michael van der Laan [ah], Domenico Ventrella [t], Michelle Viswanathan [at], Xu Xu [an], Wang Zhou [ai]

[a] Department of Bioresource Engineering, McGill University, Sainte-Anne-de-Bellevue, Quebec H9X 3V9, Canada
[b] USDA-ARS Rangeland Resources and Systems Research Unit, Fort Collins, CO 80526, USA
[c] Department of Agricultural & Biosystems Engineering, Iowa State University, Ames, IA 50011-1098, USA
[d] Ottawa Research and Development Centre, Agriculture & Agri-Food Canada, Ottawa, Ontario K1A 0C6, Canada
[e] Harrow Research and Development Centre, Agriculture and Agri-Food Canada, Harrow, ON N0R 1G0, Canada
[f] Faculty of Organic Agricultural Sciences, University of Kassel, Germany
[g] USDA-ARS Agroecosystem Management Research Unit, Lincoln, NE 68583-0937, USA
[h] Department of Biosystems Engineering, Zhejiang University, 866 Yuhangtang Road, Hangzhou, Zhejiang 310058, China

* Corresponding author.
*E-mail addresses:* viveka.nand@mail.mcgill.ca (V. Nand), zhiming.qi@mcgill.ca (Z. Qi), liwang.ma@usda.gov (L. Ma), mhelmers@iastate.edu (M.J. Helmers), chandra.madramootoo@mcgill.ca (C.A. Madramootoo), ward.smith@agr.gc.ca (W.N. Smith), Tiequan.Zhang@agr.gc.ca (T. Zhang), tobias.weber@uni-kassel.de (T.K.D. Weber), elizabeth.pattey@agr.gc.ca (E. Pattey), leo.li@mail.mcgill.ca (Z. Li), jiaxin.wang3@mail.mcgill.ca (J. Wang), virginia.jin@usda.gov (V.L. Jin), jqj713@zju.edu.cn (Q. Jiang), mario.tenuta@umanitoba.ca (M. Tenuta), thomas.trout@ars.usda.gov (T.J. Trout), yzchhm@yzu.edu.cn (H. Cheng), daren.harmel@usda.gov (R.D. Harmel), bruce.kimball@usda.gov (B.A. Kimball), kelly.thorp@usda.gov (K.R. Thorp), kjboote@ufl.edu (K.J. Boote), stockle@wsu.edu (C. Stockle), asuyker1@unl.edu (A.E. Suyker), Steve.Evett@usda.gov (S.R. Evett), david.brauer@usda.gov (D.K. Brauer), gwen.coyle@usda.gov (G.G. Coyle), karen.copeland@usda.gov (K.S. Copeland), gary.marek@usda.gov (G.W. Marek), paul.colaizzi@usda.gov (P.D. Colaizzi), marco.acutis@unimi.it (M. Acutis), m.alimagham@gmail.com (S.M. Alimagham), sarchont@iastate.edu (S. Archontoulis), babacar.faye@ird.fr (F. Babacar), zoltan.barcza@ttk.elte.hu (Z. Barcza), basso@msu.edu (B. Basso), patrick.bertuzzi@inra.fr (P. Bertuzzi), julie.constantin@toulouse.inra.fr (J. Constantin), Max.DeAntoni@des.qld.gov.au (M. De Antoni Migliorati), Benjamin.Dumont@uliege.be (B. Dumont), jean-louis.durand@inra.fr (J.-L. Durand), fodor.nandor@atk.hu (N. Fodor), tgaiser@uni-bonn.de (T. Gaiser), pasquale.garofalo@crea.gov.it (P. Garofalo), sebastian.gayler@uni-hohenheim.de (S. Gayler), luisa.giglio@crea.gov.it (L. Giglio), rgrant@ualberta.ca (R. Grant), kaiyug@illinois.edu (K. Guan), gerrit@ufl.edu (G. Hoogenboom), soohkim@uw.edu (S.-H. Kim), ikisekka@ucdavis.edu (I. Kisekka), jon.lizaso@upm.es (J. Lizaso), sara.masia@cmcc.it (S. Masia), S20193091624@cau.edu.cn (H. Meng), valentina.mereu@cmcc.it (V. Mereu), mukhtar.ahmed@slu.se (A. Mukhtar), alessia.perego@unimi.it (A. Perego), binpeng@illinois.edu (B. Peng), priesack@helmholtz-muenchen.de (E. Priesack), vakhtang.shelia@ufl.edu (V. Shelia), rlsnyder@ucdavis.edu (R. Snyder), spano@uniss.it (D. Spano), amit.srivastava@uni-bonn.de (A. Srivastava), rgrant@ualberta.ca (A. Thomson), Dennis.Timlin@ars.usda.gov (D. Timlin), antonio.trabucco@cmcc.it (A. Trabucco), webber@zalf.de (H. Webber), magali.willaume@ensat.fr (M. Willaume), karina.williams@metoffice.gov.uk (K. Williams), ah.michael.vanderlaan@up.ac.za (M. van der Laan), domenico.ventrella@crea.gov.it (D. Ventrella), u16015925@tuks.co.za (M. Viswanathan), xushengwu@cau.edu.cn (X. Xu), wangzhou@illinois.edu (W. Zhou).

https://doi.org/10.1016/j.jhydrol.2025.133631
Received 5 September 2024; Received in revised form 8 April 2025; Accepted 30 May 2025
Available online 3 June 2025

[i] Faculty of Agricultural and Food Sciences, University of Manitoba, Canada
[j] USDA-ARS, Water Management Unit, Fort Collins, CO 80526, USA
[k] School of Environmental Science and Engineering, School of Hydraulic Science and Engineering, Yangzhou University, Yangzhou 225127, China
[l] USDA-ARS, Center for Agricultural Resources Research, Fort Collins, CO 80526, USA
[m] U.S. Arid-Land Agricultural Research Center, USDA-ARS, Maricopa, AZ 85138, USA
[n] University of Florida, Agricultural and Biological Engineering, Frazier Rogers Hall, Gainesville, FL 32611-0570, USA
[o] Biological Systems Engineering, Washington State University, 1935 E. Grimes Way, PO Box 646120, Pullman, WA 99164-6120, USA
[p] School of Natural Resources, University of Nebraska-Lincoln, Lincoln, NE, USA
[q] Conservation and Production Research Laboratory, USDA-ARS, Bushland, TX, USA
[r] Department of Agricultural and Environmental Sciences, University of Milan, Via Celoria 2, 20133 Milan, Italy
[s] Agronomy Group, Gorgan University of Agricultural Science and Natural Resources, Gorgan 49138-15739, Iran
[t] Iowa State University, Department of Agronomy, Ames, IA 50010, USA
[u] Institut de recherche pour le développement (IRD) ESPACE-DEV, F-34093 Montpellier Cedex, France
[v] ELTE Eötvös Loránd University, Department of Meteorology, H-1192 Budapest, Hungary
[w] Czech University of Life Sciences Prague, Faculty of Forestry and Wood Sciences, 165 21 Prague, Czech Republic
[x] Michigan State University, Department of Geological Sciences, W.K. Kellogg Biological Station, 288 Farm Ln, 307 Natural Science Bldg., East Lansing, MI 48824, USA
[y] US1116 AgroClim, INRAE centre de recherche Provence-Alpes-Côte d'Azur, 228, route de l'Aérodrome, CS 40 509, Domaine Saint Paul, Site Agroparc, 84914 Avignon Cedex 9, France
[z] AGIR, Université de Toulouse, INRAE, INPT, INP-EI PURPAN, 24 Chemin de Borde Rouge – Auzeville CS, 52627 CastanetTolosan, France
[aa] Queensland Department of Environment and Science, Queensland, Australia
[ab] ULiege-GxABT, University of Liege – Gembloux Agro-Bio Tech, TERRA Teaching and Research Centre, Plant Science Axis/Crop Science Lab, B-5030 Gembloux, Belgium
[ac] Unité de Recherches Pluridisciplinaire Prairies et Plantes Fourragères, INRAE, 86 600 Lusignan, France
[ad] Agricultural Institute, Centre for Agricultural Research, H-2462 Martonvásár, Brunszvik u. 2., Hungary
[ae] Institute of Crop Science and Resource Conservation, University of Bonn, Katzenburgweg, 5D-53115 Bonn, Germany
[af] Council for Agricultural Research and Economics, Agriculture and Environment Research Center, CREA-AA, Via Celso Ulpiani 5, 70125 Bari, BA, Italy
[ag] Universitat Hohenheim, Institute of Soil Science and Land Evaluation, Biogeophysics, Emil-Wolff-Str. 27, D-70593 Stuttgart, Germany
[ah] Department of Renewable Resources, University of Alberta, Edmonton, Alberta T6G 2E3, Canada
[ai] College of Agricultural, Consumer and Environmental Sciences (ACES), University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA
[aj] School of Environmental and Forest Sciences, University of Washington, Center for Urban Horticulture, Seattle, WA 98195, USA
[ak] Agricultural Water Management and Irrigation Engineering, University of California Davis; Agricultural Water Management and Irrigation Engineering, University of California Davis; Departments of Land, Air, and Water Resources and of Biological and Agricultural Engineering, One Shields Avenue, PES 1110, Davis, CA 95616-5270, USA
[al] Technical University of Madrid (UPM), Dept. Producción Agraria-CEIGRAM, Ciudad Universitaria, 28040 Madrid, Spain
[am] Land and Water Management Department, IHE Delft Institute for Water Education, Delft, the Netherlands
[an] China Agricultural University, Beijing, China
[ao] CMCC Foundation-Euro-Mediterranean Centre on Climate Change, Lecce, Italy
[ap] Department of Agronomy, PMAS Arid Agriculture University, Rawalpindi, Pakistan
[aq] Swedish University of Agricultural Sciences, Umea, Sweden
[ar] Helmholtz Center Munich, Institute of Biochemical Plant Pathology, Ingolstaedter Landstr, 185764 Neuherberg, Germany
[as] University of California Davis, USA
[at] University of Pretoria, Pretoria, South Africa
[au] Crop Systems and Global Change Research Unit, USDA-ARS, Beltsville, MD, USA
[av] Leibniz Centre for Agricultural Landscape Research (ZALF), Mucheberg 15374, Germany
[aw] Met Office Hadley Centre, FitzRoy, Road, Exeter, Devon EX1 3PB, United Kingdom

## ARTICLE INFO

This manuscript was handled by Yuefei Huang, Editor-in-Chief, with the assistance of Zailin Huo, Associate Editor

*Keywords:*
Maize
Multiple crop models
Evapotranspiration
Yield
Multi-model averaging approaches

## ABSTRACT

Combining multi-model simulations can reduce the uncertainty in model structure and increase the accuracy of agricultural systems modeling results. This improvement is essential for supporting better decision making in irrigation planning and climate change adaptation strategies. Besides the commonly used arithmetic mean and median, many multi-model averaging approaches (MAA), widely examined in groundwater and hydrological modeling, but these additional MAA have not been examined in agricultural system modeling to improve the simulation accuracy. Therefore, the objective of this study is to evaluate the performance of seven MAA: two equal weighted approaches (Simple Model Averaging (SMA) and Median) and five weighted approaches (Inverse Ranking (IR), Bates and Granger Averaging (BGA), and Granger Ramanathan A, B, and C (GRA, GRB, and GRC)) in combining results of multiple agricultural system models. The Granger Ramanathan methods differ in their constraints: GRA employs conventional least squares, GRB requires non-negative weights that total to one, and GRC reduces absolute errors for robustness against outliers. The evaluation was conducted using maize yield and daily ETa simulations for both blind (uncalibrated) and calibrated phases of data from two groups of maize sites (Group A and Group B) across North America. The modeling results from the blind and calibrated phases were combined for all maize models and group maize models. Overall, all MAA performed better than individual crop models for blind and calibration phases. Specifically, the GRB model averaging method provided the closest match to measured values for daily ETa, while GRA was the most accurate for maize yield in most cases across all sites and phases. GRB improved daily ETa estimation over the median by an average of 4 % and 8.5 % in terms of RRMSE, while GRA enhanced maize yield estimation over the median by 7.5 % and 10.9 % for Group A and Group B sites, respectively. Notably, the improvement was greater in the blind phase for both groups of maize sites. An ensemble of group maize models with varied structures performed nearly as well as an ensemble of all maize models in simulating daily ETa and yield for Group A and Group B sites. Based on the results, we recommend GRA for crop yield and GRB for ETa simulations for maize, but both methods require observed yield and ETa data for their application; however, in the absence of observed data, we recommend the SMA method as

it performs better than the median. However, the performance of these MAA methods may differ for other crops (e.g., soybean, wheat, canola, potato, alfalfa) or regions, and it should be evaluated in future studies.

## 1. Introduction

Accurate prediction of crop yield and actual crop evapotranspiration (ETa) is essential for managing water resources and optimizing crop production in agriculturally dominated regions. These predictions are crucial for farmers, policymakers, and researchers to develop sustainable crop management strategies to mitigate the impacts of natural disasters and climate change. Agricultural system models are used to simulate the crop yield and ETa under different climate conditions, soil type, and management practices (Motha, 2011; Deb et al., 2022). These models play a pivotal role in understanding how crops respond to different climatic conditions and crop management practices. Over the years, numerous crop models, ranging from simple to complex, have been developed to simulate crop yield and ETa for different crops (Kimball et al., 2023). However, multi-crop models inter comparison studies show that no single crop model consistently outperforms others across different climate conditions due to potential issues with model structure, parameters, input data, and calibration data (Bassu et al., 2014; Fang et al., 2019). For example, the study by Bassu et al. (2014) revealed that simulated maize yields ranged from 10 to 12.5 Mg/ha in Lusignan (France), 8.5 to 12 Mg/ha in Ames (USA), 6 to 8 Mg/ha in Rio Verde (Brazil), and 4.5 to 6 Mg/ha in Morogoro (Tanzania) across 17 calibrated maize models. In another study, Kothari et al. (2022) used ten soybean models to simulate soybean yield at Azul, Argentina (ARGN); Brasilia, Brazil (BRZL); Auzeville, France (FRNC); and Ames, IA (IOWA) and Fayetteville, Arkansas (AKNS), USA, and found that the performance of DSSAT was superior at Fayetteville, DNDC at Azul and Brasilia, MONICA at Auzeville, SSM at Ames. Similar variability in simulated maize yield and daily and seasonal ETa simulations were noted by Kimball et al. (2019), indicating the challenges of precisely simulating the yield and ETa. These variations in ETa and yield predictions can raise the question which model should be used for precisely simulating crop yield and ETa across diverse climatic conditions (Martre et al., 2015; Kothari et al., 2022; Kimball et al., 2023).

These challenges are notably crucial in regions where precise predictions of ETa are critical for irrigation scheduling and water resource management. Therefore, there is a need for reliable methods that can improve the simulation precision of crop model predictions across various climatic regions. Studies on crop modeling have shown that an ensemble of output of multiple crop models is more reliable and efficient than individual models (Bassu et al., 2014; Kothari et al., 2022; Kimball et al., 2023). Multiple crop model ensembles reduce errors by achieving an optimal balance between bias and variance. In Agricultural Model Intercomparison and Improvement Project (AgMIP), studies, the estimated mean and median of multiple crop models outputs (yield and ETa), demonstrated better simulation accuracy than single crop models. Both approaches give equal weightage to all models without considering the performance of the models. Weighted MAA is an alternative approach which combine outputs from multiple models, by assigning weights based on each model's performance, increasing the accuracy of ensemble predictions than mean and median. While weighted ensemble predictors have been widely used in hydrological, groundwater and weather forecasting modeling, and found better results than simple mean and median methods (Ajami et al., 2006; Arsenault et al., 2015; Kumar et al., 2015; Jafarzadeh et al., 2022; Wan et al., 2021; Wallach et al., 2016). Arsenault et al. (2015) compared nine MAA across 429 catchments and found that the Granger Ramanathan C (GRC) method was best to combine the stream flow than others. Similarly, Kumar et al. (2015) evaluated ten different MAA methods and concluded that Granger Ramanathan B (GRB) was the most suitable MAA method to ensemble the river discharge.

The application of weighted MAA in crop modeling has not received much attention. A few studies demonstrated better results than the mean and median when they ensemble simulations using Bayesian model averaging (BMA) (Neuman, 2003; Huang et al., 2017; Gao et al., 2021). Numerous other weighted MAA, such as inverse rank, multiple linear regression (Kumar et al., 2015), machine learning algorithms (Zaherpour et al., 2019), and Information Criterion Averaging (Akaike, 1974; Schwarz, 1978), are also discussed in the literature and widely used in hydrological and groundwater modeling studies. But they are rarely applied in crop modeling. Therefore, there is an opportunity to explore other weighted MAA methods for increasing the simulation accuracy of crop yield and ETa across diverse climate, soil and management conditions.

Crop yield and ETa simulation accuracy can be increased by calibrating crop model parameters using various observed data sources. These include field experimental data, such as initial water content, phenological events, soil water content, leaf area index (LAI), daily $ET_a$, biomass, and yield. However, these measured data sets are often not available at many sites, and the limited availability of measured data can remarkably impact the predictive capabilities of individual crop models in predicting crop yields and $ET_a$. In past AgMIP maize modeling studies, the mean or median of yield and daily ETa simulations were satisfactory under blind phase (uncalibrated) and calibrated phase. However, there is a need to examine whether weighted MAA can further improve the simulation accuracy for different climatic conditions.

The purpose of this study is to address the aforementioned research gaps. The effectiveness of seven MAA techniques to ensemble daily ETa and maize yield simulations during both blind and calibrated phases was assessed. The study also determined the best MAA technique for varied soil, climate and management conditions in the United States and Canada. There were elven maize field experiments sites selected across the USA and Canada. We divided all sites into Group A and Group B. Five models were used to simulate maize yield and ETa at Group A sites (nine sites) which falls in USA and Canada. For Group B sites (Mead, N and Bushland, Tx), ETa and yield simulations of 41 maize models were used from a previous AgMIP study (Kimball et al., 2023).

## 2. Materials and methods

### 2.1. Description of field experiment sites and experiment data

Nine maize (*Zea mays* L.) field experiment sites (Group A) were selected for analysis: Ames (Iowa, USA), Gilmore (Iowa, USA), Greeley (Colorado, USA), Ithaca (Nebraska, USA), Glenlea (Manitoba, Canada), Harrow (Ontario, Canada), Ottawa (Ontario, Canada), Sainte-Anne-de-Bellevue (Quebec, Canada), and Saint Emmanuel (Quebec, Canada) (Table 1 and Fig. 1). In addition, two maize field sites (Group B) previously used for AgMIP maize project ETa and yield simulations studies (Mead and Bushland) were selected, focusing on four treatments (i.e., Mead rainfed, Mead irrigated, Bushland 75 % Mid Elevation Sprinkler Application (MESA) irrigation, Bushland 100 % MESA irrigation). The Bushland, Mead, Ithaca, and Greeley sites were irrigated while the remaining sites were rainfed. The average growing season air temperature, rainfall, and soil types of each site are given in Table 1. The average growing season temperature varied between 10.40 °C in Ithaca, USA, and 22.80 °C in Bushland, USA, while seasonal precipitation ranged from 191 mm in Greeley, USA, to 592.36 mm in Ithaca, USA across the maize experiment sites. Data availability period of each site is given in Table 1. A detailed description of available measurements of each site is given in Supplementary information Table S1. In-situ measured daily weather data, including maximum and minimum air

temperature, rainfall, wind speed, relative humidity, and solar radiation, were utilized for all sites except Sainte-Anne-de-Bellevue, where specific site weather data were not measured. Weather data for Sainte-Anne-de-Bellevue was obtained from the nearest weather station of Environment Canada. For soil-related information, measured soil profile data were used across all sites. Comprehensive crop management details, including tillage practices, cultivar details, seeding rate, seeding date, plant density, fertilizer application rate, harvesting date, biomass, and grain yield were obtained for all sites. The quantity and timing of irrigation was obtained for the irrigated sites. Phenological dates, detailing the various stages of plant development, were meticulously recorded for Ames, Bushland, Greeley, Mead, Ottawa, and Saint Emanuel. Additionally, time-series measurements of Leaf Area Index (LAI) and actual crop evapotranspiration (ETa) were obtained for Ames, Bushland, Greeley, Mead, and Ottawa. Measured layer-wise soil water content data were available for all sites except Harrow and Sainte-Anne-De-Bellevue.

### 2.2. Crop model setup and calibration

As mentioned in Section 2.1, we used crop yield and ETa simulations from several field experiment sites. These field experiment sites were divided into two groups i.e. Group A and Group B. Group A sites was comprised of simulated crop yield and ETa data from the uncalibrated (Blind Phase) and fully calibrated phases of the five maize models in this study (Table 1 and Table S2). Group B sites included simulated daily ETa and yield data from uncalibrated and fully calibrated phases of 41 maize models for the Bushland and Mead sites. This data was sourced from AgMIP maize study (Kimball et al., 2023). The description of 41 Maize Models is given in Supplementary information Table S3. A detailed explanation of the model set-up and calibration process is presented in Kimball et al. (2023).

In the present study, for Group A sites, five Maize models were selected from the top seven fully calibrated maize models identified in the AgMIP Maize study (Kimball et al., 2019). These maize models include DSSAT-CERES maize with Priestly-Taylor Ritchie ET equation (DCPR), DSSAT-CERES maize with FAO56 Ritchie ET equation (DCFR), APSIM-maize with SOILWAT Archontoulis subroutine (AMW), APSIM-maize with SWIM Archontoulis subroutine (AMSA), and RZWQM2 (Table S2).The selection of five maize models was based on their performance to simulate growing season daily ETa, maximum LAI, biomass and grain yield over the eight years growing season. All these maize models were ranked among the top seven maize models to simulate the same over the study period. The RZWQM2 model which uses the Shuttleworth-Wallace approach to estimate potential transpiration (PT) and potential evaporation (PE) (Shuttleworth and Wallace, 1985) did not perform well in simulating ETa among the top seven maize models, however, it was in the top seven maize models' performer in simulating maximum LAI, biomass and crop yield and therefore it was included in this study.

All these five maize models were used to simulate crop yield and ETa for Group A's sites (Table S2). Maize models were calibrated and validated using measured field data (Kimball et al., 2019). Models were set up utilizing site-specific measured data, encompassing layered soil texture along with corresponding physical and hydraulic properties, tillage dates, cultivar details, seeding dates, plant density, irrigation amounts, and fertilizer rates.

In the blind phase (uncalibrated phase), for Group A sites, all five maize models were set up using site-specific measured input data, including soil, weather, and crop management details (such as seeding date, plant density, and fertilizer rate). The models' phenology parameters were then adjusted to align with the crop maturity dates across all sites. Subsequently, the models were run to simulate ETa and yield. During this phase, models were not calibrated with available LAI, soil moisture, ETa, and yield data.

In the calibrated phase, all maize models were fine-tuned against the measured data to improve their ETa and crop yield simulation accuracy. We followed the step-by-step calibration procedure given in AgMIP maize study (Kimball et al., 2019). Cultivar parameters in each model were initially adjusted to align anthesis, silking, and maturity dates with observed ones depending on sites and available phenological measurement dates. Then, maize models were calibrated for LAI. Subsequently,

**Table 1**
Details of selected crop field sites and corresponding soil type, average rainfall, and average temperature during the growing season (April–October).

| Name | Country | Province State | Lat | Long | Soil type | Growing season climatic parameters | | Modeled component | Data availability period | Sources |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | Rainfall (mm) | Mean temp (°C) | | | |
| | **Group A sites** | | | | | | | | | |
| Ames | USA | Iowa | 42.02 | −93.75 | Loam | 536.37 | 18.62 | Yield and ETa | 2006–2013 | Kimbal et al., 2019 |
| Gilmore | USA | Iowa | 42.73 | −94.45 | Clay Loam | 559.35 | 17.47 | Yield | 2005–2009 | Qi et al.,2011 |
| Glenlea | Canada | Manitoba | 49.64 | −97.16 | Clay | 399.00 | 14.10 | Yield | 2006–2012 | Uzoma et. al., 2015 |
| Greeley | USA | Colorado | 40.44 | −104.00 | Loamy Sand | 191.00 | 16.50 | Yield and ETa | 2008–2013 | Qi et al.,2016 |
| Harrow | Canada | Ontario | 42.22 | −82.73 | Clay Loam | 505.93 | 18.21 | Yield | 2008–2011 | Jiang et al.,2020 |
| Ithaca | USA | Nebraska | 41.16 | −96.41 | Silty Loam | 592.36 | 10.40 | Yield | 2001–2015 | Cheng et al., 2021 |
| Ottawa | Canada | Ontario | 45.38 | −75.72 | Loam | 530.80 | 16.19 | Yield and ETa | 2002–2018 | Crépeau et al.,2021 |
| St. Emmanuel | Canada | Québec | 45.32 | −74.17 | Clay Loam | 578.87 | 16.35 | Yield | 2005–2013 | Singh, 2013 |
| Ste.-Anne-de-Bellevue | Canada | Québec | 45.43 | −73.93 | Loamy Sand | 580.52 | 16.27 | Yield | 2008–2009 | Jiang et al., 2022 |
| | **Group B Sites** | | | | | | | | | |
| Bushland | USA | Texas | 35.18 | −102.09 | Silty Clay | 350 | 22.80 | Yield and ETa | 2013,2016 | Kimbal et al., 2023 |
| Mead Rainfed | USA | Nebraska | 41.17 | −96.43 | Silty Loam | 592 | 19.90 | Yield and ETa | 2003–2013 | Kimbal et al., 2023 |
| Mead Irrigated | USA | Nebraska | 41.16 | −96.47 | Silty Loam | 592 | 19.90 | Yield and ETa | 2003–2013 | Kimbal et al., 2023 |

the models calibrated against soil water content data by adjusting saturated and lateral hydraulic conductivity depending on maize models for all sites except Harrow and Sainte-Anne-de-Bellevue. Following this, the models were fine-tuned for ETa by adjusting parameters related to albedo, soil resistance, and leaf stomatal resistance depending on specific maize model at sites (Ames, Ottawa and Greeley) those had ETa measurements. Lastly, the models were calibrated for crop yield by adjusting cultivar parameters influential on crop yield. Among the field experiment sites, maize models were comprehensively calibrated for growth stage dates, LAI, soil water content, ETa and yield for Greeley, Ames, and Ottawa sites. For the remaining sites, calibration was limited to growth stage dates, LAI, soil water content, and yield. We did not calibrate the maize models for daily ETa for remaining sites as daily ETa measurements were not available.

The calibration procedure for Group B sites is described in Kimball et al. (2023). In the present study, ETa and yield simulations from blind phase and full calibrated phase of 41 maize models were used. In the blind phase, cultivar parameters were fine tuned to match with measured anthesis, silking and maturity dates at all sites. Next, models were calibrated for LAI and biomass data using measured LAI and biomass data. Then, soil water content and ETa were calibrated by adjusting albedo, soil resistance, and stomatal resistance depending on the specific maize model. At the end, maize models were tuned to match the observed yield.

### 2.3. Model averaging approaches (MAA)

The simulated yield and daily ETa from all sites were ensembled using seven MAA: Simple Model Averaging (SMA), Median, Inverse Rank (IR), Bates and Granger Averaging (BGA), and three variants of Granger Ramanathan (GRA, GRB, and GRC) (Supplementary information Table S4). We selected simple mean and median MMA because they are widely used in agricultural system modeling and do not require measured data to estimate the weight of each model. Both methods can be applied in data-scarce regions. Weighted-based MMA, such as Bates and Granger Averaging (BGA) and Inverse Rank (IR), were selected because their ensemble performed better than calibrated models in previous studies (Aiolfi and Timmermann, 2006; Arsenault et al., 2015; Wan et al., 2021). Granger Ramanathan A, B, and C (GRA, GRB, and GRC) selected based on previous studies as their performance was similar or better than advanced MAA such as Bayesian Model Averaging (BMA) and Mallows Model Averaging (MAAS) (Diks and Vrugt, 2010; Arsenault et al., 2015; Wan et al., 2021). GRA, GRB, and GRC are less computationally expensive than BMA and MAAS.

The weight of each crop model for yield and ETa was estimated using seven MAA, which are built into the Geometric Forecast Combination (GeomComb) R package (https://github.com/cran/GeomComb). GRA, GRB and GRC MAA are represented as Ordinary Least Squares Forecast Combination (comb_OLS), Constrained Least Squares Forecast Combination (comb_CLS) and Least Absolute Deviation Forecast Combination (comb_LAD), respectively, in the GeomComb R package.

The ensemble yield and daily ETa were determined by multiplying the weight of each maize model by its corresponding simulated yield and daily ETa for each site. First, the simulated yield and daily ETa from all selected maize models were combined. Those were five for Group A sites, and 41 for Group B sites. We referred to as “all maize models.” Next, the simulated yield and daily ETa of one representative model from each model family were selected and ensembled. It was referred to as “group maize models”. The selection was based on the over all performance of models to simulate yield and daily ETa within each family for calibrated and un-calibrated phase. If a model family had no variants, it was selected by default.

For Group A sites, three group maize models were selected, while for Group B sites, twenty-two group maize models were chosen. Selected Group maize models are given in the Supplementary Tables S2 and S3.

The simulated yield was ensemble across all sites, while the simulated daily ETa was ensemble for three sites in Group A (Ames, Greeley, and Ottawa) and all sites in Group B. The resulting yields and daily ETa obtained through the MAA methods were subsequently compared with the observed yield and daily ETa datasets. Details of the multiple MAA are given below:

a. **Simple Model Averaging (SMA):** In this approach, the weight of each model is assigned equally. Mathematically, it can be estimated as:

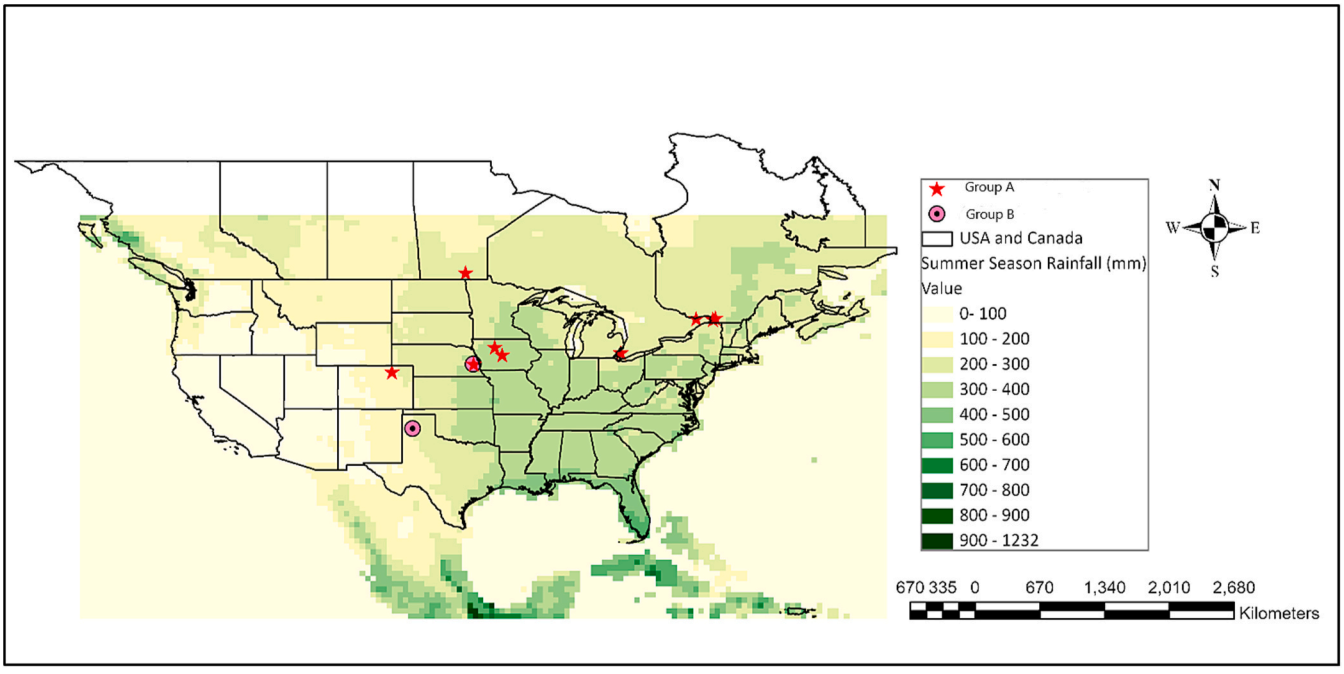



**Fig. 1.** Locations of crop field sites in the USA and Canada (Group A sites, and Group B sites).
Source: http://drought.memphis.edu/naspa/CompReconRange.aspx

$$W = \frac{1}{n} \quad (1)$$

where n is the number of ensemble models, and W is the estimated weight of each ensemble model.

b. **Median:** The median of simulated values of all ensemble models is taken to combine the forecast.
c. **Inverse rank:** The inverse rank approach, rank each ensemble model based on their simulation performance. The first rank is assigned to model with lowest root mean squared error, the model with the second lowest mean squared error is assigned the rank 2. Then weightage of each model is calculated as follows:

$$W = \frac{Rank_i^{-1}}{\sum_{i=1}^{N} Rank_i^{-1}} \quad (2)$$

where W is the estimated weight of each ensemble model. $Rank_i$ is the rank of the $i_{th}$ ensemble model.

d. **Bates and Granger Averaging (BGA):** The BGA method combined the forecast of ensemble models by minimizing the root mean squared error between simulated and observed values. It can be estimated as:

$$W = \frac{\frac{1}{RMSE_i^2}}{\sum_{i=1}^{N} \frac{1}{RMSE_i^2}} \quad (3)$$

where W is the estimated weight of each ensemble model. $RMSE_i$ is the root mean square error of the $i^{th}$ ensemble model.

e. **Granger Ramanathan A (GRA):** The GRA approach, developed by Granger and Ramanathan in 1984, employs the ordinary least squares (OLS) method to estimate weight of each model, effectively lowering the sum of squared error (SSE) but lacking bias correction. Weight of each ensemble model are estimated by following equation:

$$W = \left(ETsim^T ETsim\right)^{-1} ET^T{}_{sim} ET_{meas} \quad (4)$$

where $ET_{sim}$ is the matrix of the maize models' simulations, $ET_{meas}$ is the matrix of measured values, and $ETsim^T$ is the transpose matrix of the maize models' simulations.

f. **Granger Ramanathan B (GRB):** GRB uses constrained least squares (CLS) method, ensuring that the weights of all models sum to one. In GRB, weights are estimated by:

$$W = \left(ETsim^T ETsim\right)^{-1} ET^T{}_{sim} ET_{meas} - \lambda_B \left(ETsim^T ETsim\right)^{-1} l \quad (5)$$

$$\lambda_B = \frac{\left(l^T (ETsim^T ETsim)^{-1} ET^T{}_{sim} ET_{meas} - 1\right)}{\left(l^T (ETsim^T ETsim)^{-1} l\right)} \quad (6)$$

where $\lambda_B$ is a Lagrangian multiplier, l is the unit vector of same dimension as that vector of W.

g. **Granger Ramanathan C (GRC):** The GRC approach is similar to GRA but includes a bias correction term. In GRC, weights are estimated by:

$$W = \left(ETsim^T ETsim\right)^{-1} ET^T{}_{sim} ET_{meas} - \delta \left(ETsim^T ETsim\right)^{-1} ET^T{}_{sim} l \quad (7)$$

$\delta$ is a bias correction term which is estimated by following relationship:

$$\delta = \frac{l^T e_A}{(n - \theta)} \quad (8)$$

where $e_A$ is the vector of errors (ETmeas-ETsim*W) estimated by GRA method and θ is estimated by following equation:

$$\theta = l^T ETsim \left(ETsim^T ETsim\right)^{-1} ET^T{}_{sim} l \quad (9)$$

Detailed information on GRA, GRB and GRC model averaging approaches can be found in Granger and Ramanathan (1984).

### 2.4. Performance evaluation of the models

The evaluation of the crop models and model averaging methods performance was assessed by statistical indicators such as relative root mean squared error (RRMSE). Jamieson et al. (1991) concluded that RRMSE values below 10 % are "excellent", values from 10 to 20 % are "good", values from 20 to 30 % are "satisfactory", and values exceeding 30 % are "poor".

$$RRMSE = \frac{100}{\bar{o}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} (o_i - s_i)^2} \quad (9)$$

where n is the number of observed and simulated data points, $o_i$ is the observed value, $s_i$ is the model simulated value, $\bar{o}$ is the mean of observed values.

## 3. Results

### 3.1. Group A sites simulations

In this section, the simulated daily ETa and seasonal yield were examined using five maize crop models (DSPR, DSFR, AMW, AMSA, and RZWQM2) across nine sites in the USA and Canada, under both the blind and calibrated phases. Additionally, the MAA estimated daily ETa and seasonal yield results were assessed. The analysis focused on daily ETa simulations at Ames, Greeley, and Ottawa, where daily ETa measurements were available. Seasonal yield was analysed at all nine sites. For Ames, Greeley, and Ottawa, the analysis focused on the growing seasons of 2006–2008, 2010 for the Ames and Greeley and 2002, 2006 and 2010 for daily ETa simulations in Ottawa, respectively.

#### 3.1.1. Blind phase

*3.1.1.1. Crop evapotranspiration.* A wide range of daily ETa simulations was observed in the five maize models at all sites, especially in the early and end-growth stages during the blind phase (Fig. 2). The RRMSE between measured and simulated daily ETa ranged from 49.8 to 72.1 % at Ames of the growing seasons of 2006–2008, from 36.5 to 104.2 % at Greeley for the 2010 growing season, and from 40.6.5 to 83.8 % at Ottawa for the growing seasons 2002, 2006 and 2010 (Fig. 3a). In 2006 at Ames, the measured average daily ETa during the growing season was 2.5 mm, while the simulated average daily ETa ranged from 2.3 to 2.7 mm/day. Similarly, at Greeley in 2010, the measured average daily ETa was 4.4 mm, and simulated average daily ETa values ranged from 3.6 to 6.9 mm/day. In Ottawa in 2006, the measured average daily ETa was 2.3 mm, while simulated values varied between 2.2 and 3.3 mm/day.

However, ensembling the daily ETa simulations from all five maize models using seven model averaging methods improved the accuracy of daily ETa simulations based on the RRMSE (Fig. 3a and Table 2). The performance of GRA model averaging method to combine daily ETa simulations was best at the Ames and Ottawa sites, whereas GRB performed slightly better at the Greeley site. Fig. 2 indicates a closer agreement between measured and GRB ensembled daily ETa over the growing season at all sites.

When daily ETa simulations of group maize models were ensembled, the performance of model averaging methods decreased compared to the ensembling of all maize models (Table 2). Though GRA and GRC model averaging methods showed almost similar performance in combining daily ETa, GRA ensemble daily was best at the Ames and

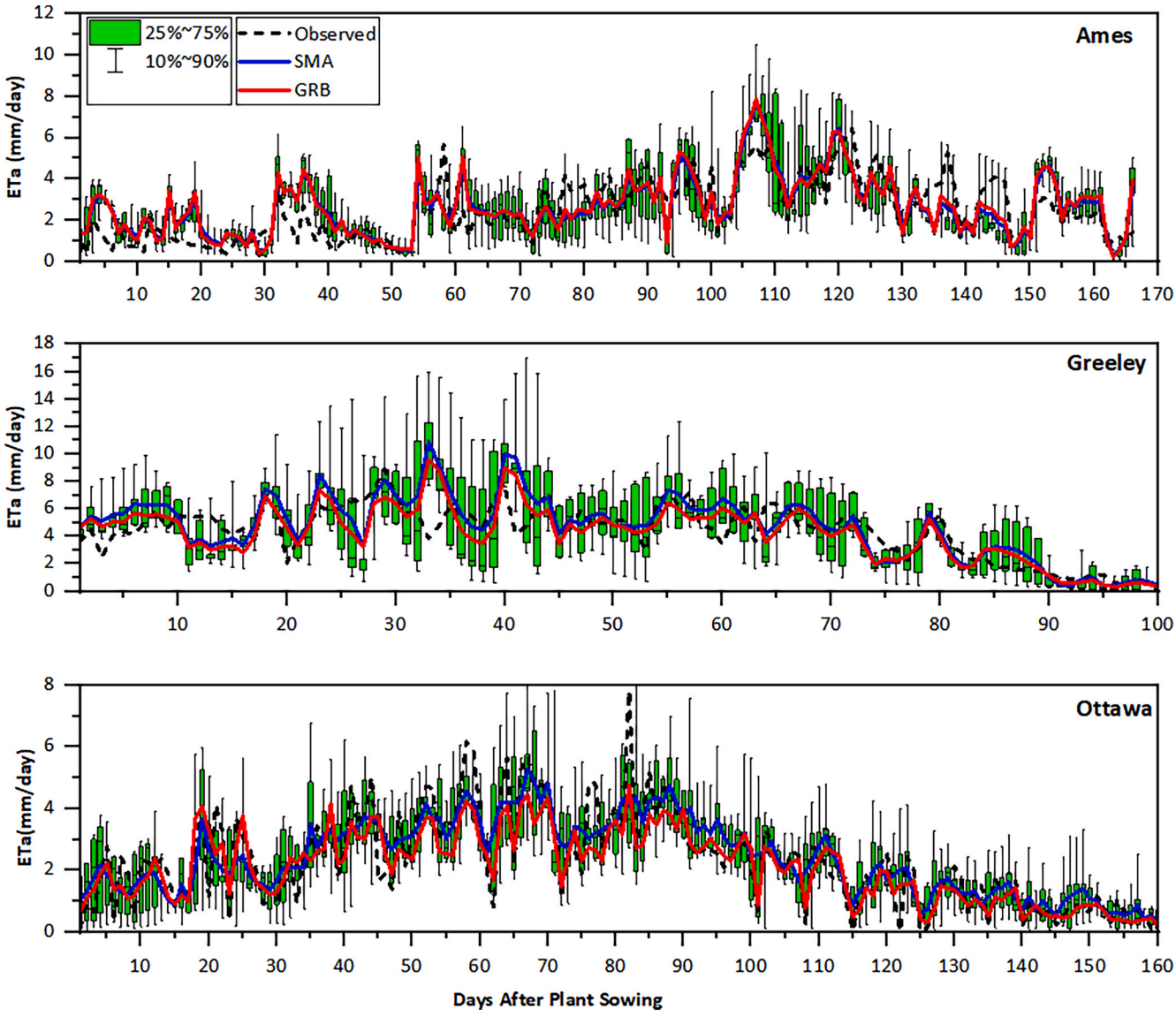



**Fig. 2.** Box plots of daily simulated evapotranspiration (ETa) across the five maize models of the maize season 2006, 2010, and 2006 at Group A sites (Ames, Greeley, and Ottawa), respectively, for the uncalibrated phase. Observed daily ETa values, and the GRB and SMA multi-model averaging approaches derived daily ETa values from the five maize models are also presented. The simulated outputs of the uncalibrated phase where all maize model were set up using in-situ data and no calibration was done.

Ottawa sites, whereas GRC performed best at the Greeley site. Overall, by taking the average of RRMSE of Ames, Greeley and Ottawa sites, the results indicate that there was slight variation noted in GRA, GRB and GRC for ensemble daily ETa. However, GRC was identified as the best model ensemble approach (Table 5).

*3.1.1.2. Crop yield.* Uncalibrated maize models showed unsatisfactory performance across all sites, as indicated by high RRMSE values (Fig. 4a). However, combining simulated yields from all maize models using model averaging methods remarkably improved yield simulation performance, achieving acceptable RRMSE criteria. Generally, the performance of GRA and GRC was similar across all sites, followed by GRB, IR, BGA, SMA, and the Median (Fig. 4). Additionally, when yield simulations from group maize models were ensembled, no improvements were found as compared to an ensemble of all maize models (Table 3). There was a slight decrease in the performance of the model averaging method in the ensemble of group maize models.

### *3.1.2. Calibrated phase*

*3.1.2.1. Crop evapotranspiration.* Moderate variability in the daily simulated ETa persisted at each site, despite calibrating all crop models (Fig. 5). The RRMSE values ranged from 41.4 to 50.8 % at Ames of the growing seasons of 2006–2008, 36.5–48.8 % at Greeley for the 2010 growing season, and 34.4–59.1 % at Ottawa for the growing seasons 2002, 2006 and 2010 (Fig. 3b), indicating that the RRMSE remained in the unacceptable range across all maize models and sites. At the Ames site, the average measured growing season daily ETa was 2.5 mm, while the average simulated daily ETa ranged from 2.6 to 2.9 mm/day across all maize models. Similarly, in Greeley, the average growing season measured daily ETa was 4.4 mm, with simulated values between 4.0 and 4.8 mm/day. Similar results were observed at the Ottawa site. However, when an ensemble of all maize models was taken using model averaging methods, this variability was reduced across all sites as shown by RRMSE values in Fig. 3b. A slightly improvement in ensembled daily ETa simulations was noted across all model averaging methods compared to the blind phase (Table 2). The RRMSE for the ensemble

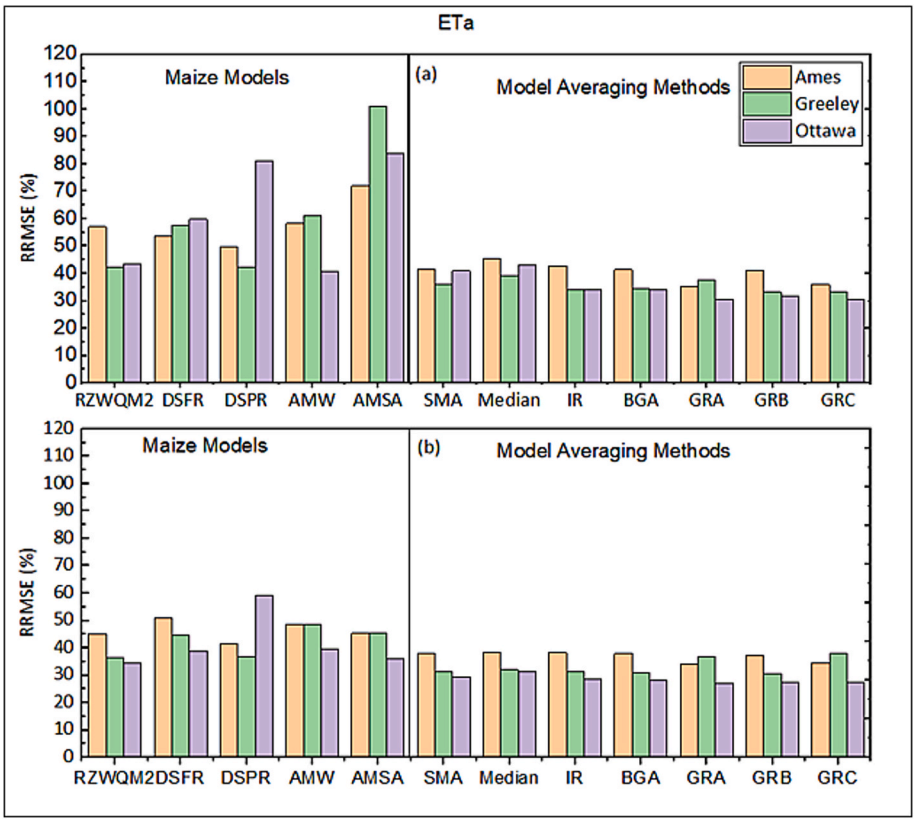



**Fig. 3.** RRMSE between the measured and simulated daily ETa across five maize models and seven multi-model averaging approaches (MAA) under uncalibrated (a) and calibrated (b) phases at Group A sites.

**Table 2**
A comparison of RRMSE between the measured daily ETa and ensembled daily ETa of all maize models and group maize models using seven multi-model averaging approaches (MAA) at Group A sites under the Blind and Calibrated Phase.

| Averaging approaches | Blind | | | | | | Calibrated | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | All models | | | Group models | | | All models | | | Group models | | |
| | Ames | Greeley | Ottawa | Ames | Greeley | Ottawa | Ames | Greeley | Ottawa | Ames | Greeley | Ottawa |
| SMA | 41.4 | 36.2 | 41.0 | 40.8 | 45.8 | 35.5 | 38.0 | 31.5 | 29.6 | 39.5 | 31.8 | 29.1 |
| Median | 45.5 | 32.8 | 43.2 | 44.5 | 45.8 | 35.8 | 38.6 | 32.0 | 31.2 | 39.8 | 33.6 | 30.5 |
| IR | 42.8 | 32.6 | 34.2 | 42.0 | 37.2 | 33.5 | 38.4 | 31.4 | 28.6 | 39.1 | 31.5 | 30.0 |
| BGA | 41.4 | 30.4 | 33.9 | 40.8 | 35.5 | 33.4 | 38.0 | 31.1 | 28.0 | 39.2 | 31.6 | 29.2 |
| GRA | 35.4 | 49.5 | 30.6 | 34.0 | 39.2 | 31.5 | 34.0 | 37.0 | 27.1 | 34.7 | 37.8 | 28.3 |
| GRB | 41.2 | 29.8 | 31.7 | 40.8 | 35.0 | 33.0 | 37.3 | 30.6 | 27.6 | 38.8 | 31.4 | 29.0 |
| GRC | 36.0 | 34.8 | 30.7 | 34.6 | 34.9 | 31.7 | 34.6 | 38.1 | 27.3 | 35.1 | 36.9 | 28.5 |

varied from 34.0 to 38.6 % at Ames, 30.6 to 38.1 % at Greeley, and 27.1 to 31.2 % at Ottawa across all MAAs over the respective goring season years. The GRA ensemble of daily ETa showed closer agreement with the measured daily ETa than other MAAs at all sites except Greeley. Furthermore, the accuracy of daily ETa improved when averaging group maize models compared to averaging all maize models (Table 2). GRA performed the best for combining daily ETa at Ames and Ottawa, while GRB was the best at the Ottawa site. The performance of MAA to ensemble daily ETa simulations was very close for all maize models and group maize models. In general, the results suggest that there was slight variation noted across all MAA for ensemble daily ETa. However, GRB was identified as the best model ensemble approach.

*3.1.2.2. Crop yield.* When all maize models were fully calibrated, their performance improved across all sites. Comparing the simulated yields of individual maize models with the measured yields, the RRMSE was found to be less than 30 % (Fig. 4b), indicating that the performance of each crop model varied depending on the site, and no single model consistently outperformed others for simulating maize yield across all locations. The RRMSE between measured and simulated yield ranged from 0.44 % to 28.90 % across all maize models and sites.

Yield simulations improved further when an ensemble of all maize models was taken using model averaging methods, as indicated by RRMSE values in Fig. 4b. The GRA produced ensembled yield values were very close to the observed yields at all sites. The performance of GRC was comparable to GRA at most sites with slight variation. In the calibrated phase, the performance of model averaging methods was slightly better than in the blind phase.

However, a minor decrease in the accuracy of yield simulations was noted when using an ensemble of group maize models with model averaging methods, indicating that the ensemble of simulated yield from group maize models did not improve the yield simulations (Table 3). Among the model averaging methods, the ensemble yields from GRA and GRB matched the measured yields at most sites.

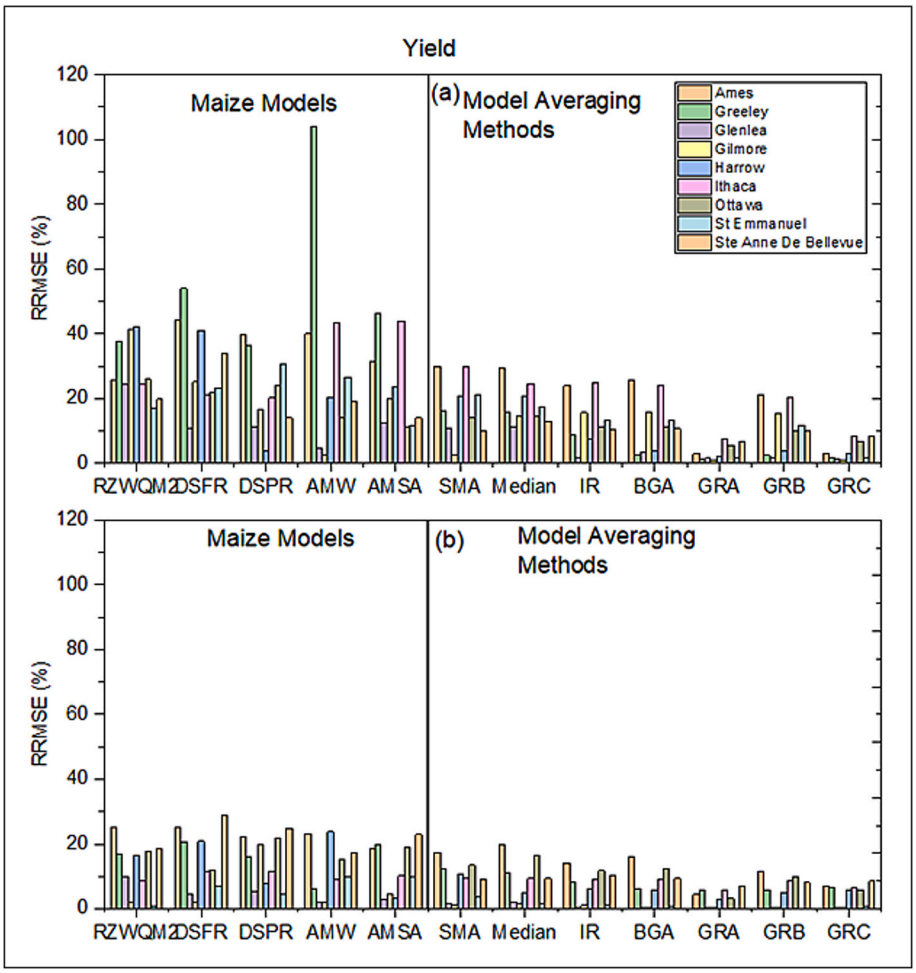



**Fig. 4.** RRMSE between the measured and simulated maize yield across five maize models and seven multi-model averaging approaches (MAA) under uncalibrated (a) and calibrated (b) phases at Group A sites.

### *3.2. Group B sites simulations*

#### *3.2.1. Blind phase*

*3.2.1.1. Crop evapotranspiration.* The 41 maize models from the AgMIP maize ET study simulated daily ETa were in a wide range at all sites (Kimball et al., 2023). The RRMSE between the daily simulated ETa and the in-situ measured daily ETa ranged from 33 % to 110 % at Mead irrigated for the growing seasons of 2003, 2005, 2007 and 2009; 32 % to 131 % at Mead rainfed for the same seasons; from 29 % to 87 % at Bushland 100 % MESA for the growing seasons of 2013 and 2016; and from 31.20 % to 79 % at Bushland 75 % MESA for the growing seasons of 2013 and 2016 across all maize models (Fig. 6a). The previous analysis by Kimball et al. (2023) revealed that the median of all maize models closely matched the measured daily ETa throughout the growing season. In the present study, variability in daily ETa simulations decreased when the ensemble of all maize models was used. Even though roughly similar performance was noted for the GRA, GRB, and GRC at all sites except Bushland 75 % MESA, overall, GRB-ensembled daily ETa performed better in matching the daily measured ETa over the growing season at most sites, followed by GRA, GRC, IR, BGA, SMA, and the Median (Table 4). The RRMSE between the ensembled daily ETa and the measured daily ETa ranged from 18.4 % to 28 % at Mead irrigated, 18.5 % to 38.1 % at Mead rainfed, 19 % to 26.4 % at Bushland 100 % MESA, and 25.8 % to 30 % at Bushland 75 % MESA sites in among MAA over the respective growing seasons (Table 4 and Fig. 6a).

The ensembled daily ETa by SMA and GRB was also compared with the measured daily ETa during the 2003 growing season at Mead's irrigated and rainfed sites. Fig. 7 illustrates a close match between the measured daily ETa and the GRB ensembled daily ETa, particularly towards the end of the growing season at the Mead Irrigated site. The GRB ensembled daily ETa followed the pattern of the measured daily ETa more closely than the SMA ensembled daily ETa. However, none of the MAAs could reproduce the peak daily measured ETa. Similarly, at the Mead rainfed site, the GRB ensembled daily ETa closely followed the daily measured ETa for the 2003 growing season (Fig. 7), whereas the SMA ensembled daily ETa showed poor agreement with the measured daily ETa, especially during the mid-and late-growing seasons. GRB ensembled daily ETa also closely followed the pattern of daily measured ETa during the 2013 crop period at Bushland 100 % MESA and 75 % MESA sites. However, the GRB and other MAA underestimated ETa during the early and mid-crop periods. This discrepancy is attributed to the inadequacy of many crop models in accounting for varying wind speed and humidity. All maize models estimated daily ETa accurately during periods of lower ETa but considerably underestimated ETa during periods of higher ETa, characterized by high wind speeds and low relative humidity (Kimball et al., 2023).

Additionally, the results of group maize models were analyzed, where one model from each crop model family was selected. This approach marginally improved the daily ETa simulations at all sites compared to considering an ensemble of all maize models (Table 4). For instance, the RRMSE between the daily measured ETa and the ensembled daily ETa of all maize models ranged from 18.4 % to 28 % across all models averaging methods at the Mead irrigated site. In contrast, the

**Table 3**
A comparison of RRMSE between the measured maize yield and ensembled maize yield of all maize models and group maize models using seven multi-model averaging approaches (MAA) at Group A sites under the Blind and Calibrated Phase.

| Averaging approaches | Blind | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | All models | | | | | | | Group models | | | | | | | |
| | Ames | Gilmore | Glenlea | Greeley | Harrow | Ithaca | Ottawa | St Emmanuel | Ste Anne De Bellevue | Ames | Gilmore | Glenlea | Greeley | Harrow | Ithaca | Ottawa | St Emmanuel | Ste Anne De Bellevue |
| SMA | 29.9 | 16.2 | 10.8 | 2.6 | 20.9 | 29.6 | 14.0 | 21.1 | 10.1 | 29.8 | 16.2 | 10.1 | 6.8 | 28.0 | 28.5 | 16.2 | 17.1 | 12.2 |
| Median | 29.2 | 15.8 | 11.1 | 14.4 | 20.8 | 24.5 | 14.4 | 17.4 | 12.9 | 31.2 | 13.3 | 9.5 | 15.5 | 37.2 | 24.5 | 16.4 | 15.7 | 13.1 |
| IR | 24.0 | 8.7 | 1.7 | 15.7 | 7.7 | 24.9 | 11.3 | 13.3 | 10.2 | 25.3 | 10.8 | 1.9 | 7.6 | 17.8 | 25.2 | 15.1 | 13.3 | 10.2 |
| BGA | 25.7 | 2.8 | 3.3 | 15.8 | 3.9 | 24.0 | 11.1 | 13.3 | 10.7 | 25.1 | 2.7 | 2.6 | 3.0 | 17.0 | 24.2 | 15.0 | 13.3 | 10.9 |
| GRA | 2.9 | 1.2 | 1.6 | 1.0 | 2.2 | 7.7 | 5.7 | 1.7 | 6.6 | 3.4 | 1.2 | 1.8 | 7.8 | 1.2 | 8.1 | 11.8 | 1.6 | 7.4 |
| GRB | 21.3 | 2.5 | 1.9 | 15.5 | 3.9 | 20.2 | 9.9 | 11.8 | 9.9 | 23.4 | 2.5 | 1.8 | 6.7 | 17.0 | 21.0 | 14.0 | 11.8 | 10.0 |
| GRC | 3.1 | 1.6 | 1.5 | 1.0 | 3.0 | 8.6 | 6.8 | 1.7 | 8.4 | 3.6 | 2.4 | 1.8 | 7.8 | 1.9 | 9.3 | 13.5 | 1.6 | 8.2 |
| | Calibrated | | | | | | | | | | | | | | | | | |
| | All models | | | | | | | Group models | | | | | | | | | | |
| SMA | 17.5 | 12.5 | 1.7 | 1.4 | 10.6 | 9.4 | 13.4 | 3.7 | 9.1 | 15.0 | 10.0 | 3.5 | 1.3 | 15.28 | 9.54 | 15.00 | 2.60 | 10.6 |
| Median | 19.6 | 10.9 | 2.0 | 1.6 | 5.0 | 9.4 | 16.3 | 1.5 | 9.3 | 13.9 | 8.7 | 4.0 | 3.2 | 20.86 | 9.28 | 17.34 | 0.72 | 12.2 |
| IR | 14.1 | 8.2 | 0.5 | 1.1 | 6.1 | 9.0 | 11.8 | 1.1 | 10.5 | 15.5 | 7.2 | 1.7 | 4.0 | 11.76 | 9.21 | 13.40 | 3.20 | 10.8 |
| BGA | 16.0 | 6.3 | 0.4 | 0.5 | 5.7 | 9.2 | 12.3 | 0.6 | 9.4 | 14.7 | 5.4 | 1.1 | 4.0 | 11.74 | 9.39 | 13.52 | 0.71 | 11.4 |
| GRA | 4.4 | 5.6 | 0.2 | 0.2 | 2.7 | 5.7 | 3.2 | 0.1 | 7.0 | 8.0 | 2.4 | 0.2 | 0.1 | 1.70 | 7.84 | 7.45 | 0.08 | 8.6 |
| GRB | 11.5 | 5.8 | 0.2 | 0.4 | 5.0 | 8.6 | 9.9 | 0.1 | 8.1 | 16.3 | 5.4 | 0.6 | 4.0 | 11.94 | 9.00 | 12.09 | 0.93 | 9.7 |
| GRC | 7.2 | 6.7 | 0.2 | 0.2 | 5.7 | 6.5 | 5.8 | 0.7 | 8.5 | 13.7 | 4.5 | 0.2 | 0.1 | 2.88 | 8.88 | 8.87 | 0.23 | 9.8 |

RRMSE between the daily measured ETa and the ensembled ETa of group maize models ranged from 18.6 % to 24.4 % across all model averaging methods. Similar findings were observed at the Mead rainfed, Bushland 100 % MESA, and Bushland 75 % MESA sites. In general, GRB ensemble approach was found best for ensemble daily ETa simulations for all maize models and group models (Table 6).

*3.2.1.2. Crop yield.* Large variability in simulated maize yields was noted across 41 maize models during the blind phase (Fig. 8a). An ensemble of simulated yields of all maize models reduced the deviation between measured yield and simulated maize yield at all sites. Among the seven MAA, GRA performed the best followed by GRC, GRB, IR, BGA, SMA, and median at most sites. Moreover, the performance of group maize models was examined. Overall, this approach improved the yield simulations for a few cases (Table 4). The performance of all MAAs in combining the simulated yield of group maize models was roughly similar to ensembling the maize yield of all maize models.

### 3.2.2. Calibrated phase

*3.2.2.1. Crop evapotranspiration.* After fully calibrating all maize models, a slight improvement in daily ETa simulations was noted in all maize models. There was still wide variability in daily ETa simulations across the 41 maize models. The RRMSE ranged from 28.5 % to 75.0 %, 30.3 % to 90.0 %, 30.0 % to 68.5 %, and 28.0 % to 67.0 % at Mead irrigated, Mead rainfed, Bushland 100 % irrigation, and Bushland 75 % irrigation sites, respectively over the corresponding growing seasons (Fig. 6b). Model averaging methods reduced the variability in daily ETa simulation by combinig daily ETa simulations of all maize models. In the calibrated phase, improvement in ensembled daily ETa simulation across MAA was slightly higher than the blind phase at all sites (Table 4). Though GRA, GRB, and GRC MAA showed almost similar performance to ensemble daily ETa of all maize models, GRA outperformed others at Mead rainfed and irrigated sites and GRB outperformed others at Bushland 75 and 100 % MESA sites. For instance, the RRMSE between the GRA ensembled daily ETa and measured daily ETa was 19.0 and 19.4 % at Mead irrigated and rainfed sites, respectively (Fig. 6b). Similarly, RRMSE between the GRB ensembled daily ETa and measured daily ETa was noted for 19.30 % and 19.40 % at Bushland 100 % MESA and 75 % MESA sites, respectively. The model averaging methods ensembled daily ETa were also compared with measured daily ETa over the growing season at Mead and Bushland sites. Fig. 9 shows a close match between in-situ measured daily ETa and GRB ensembled daily ETa, particularly during the 2003 growing season at Mead rainfed, where GRB closely followed the measured pattern.

Moreover, the ensemble of daily ETa of group maize models was compared using different model averaging methods. A slight improvement in ensembled daily ETa simulations was noted when considering group maize models (Table 4), however, the pattern of performance of MAA to ensemble daily ETa simulations of group maize models was similar to all maize models. For example, GRA model averaging method ensembled daily ETa was found best at Mead irrigated and rainfed sites, whereas GRB ensembled daily ETa outperformed to others at Bushland 100 and 75 % MESA sites in both cases (Table 4). By synthesizing results of all maize models and group maize models, GRA ensemble approach was found best for ensemble daily ETa simulations for all maize models whereas GRB ensemble approach was identified best for group maize models (Table 6).

*3.2.2.2. Crop yield.* Simulated yield showed remarkable improvement in most maize models after full calibration compared to the blind phase (Fig. 8b). The greatest improvement in yield simulations was observed at the Mead irrigated site; however, moderate variability in yield simulations was found across all maize models at the Mead rainfed, Bushland 100 % MESA, and Bushland 75 % MESA sites. This variability decreased

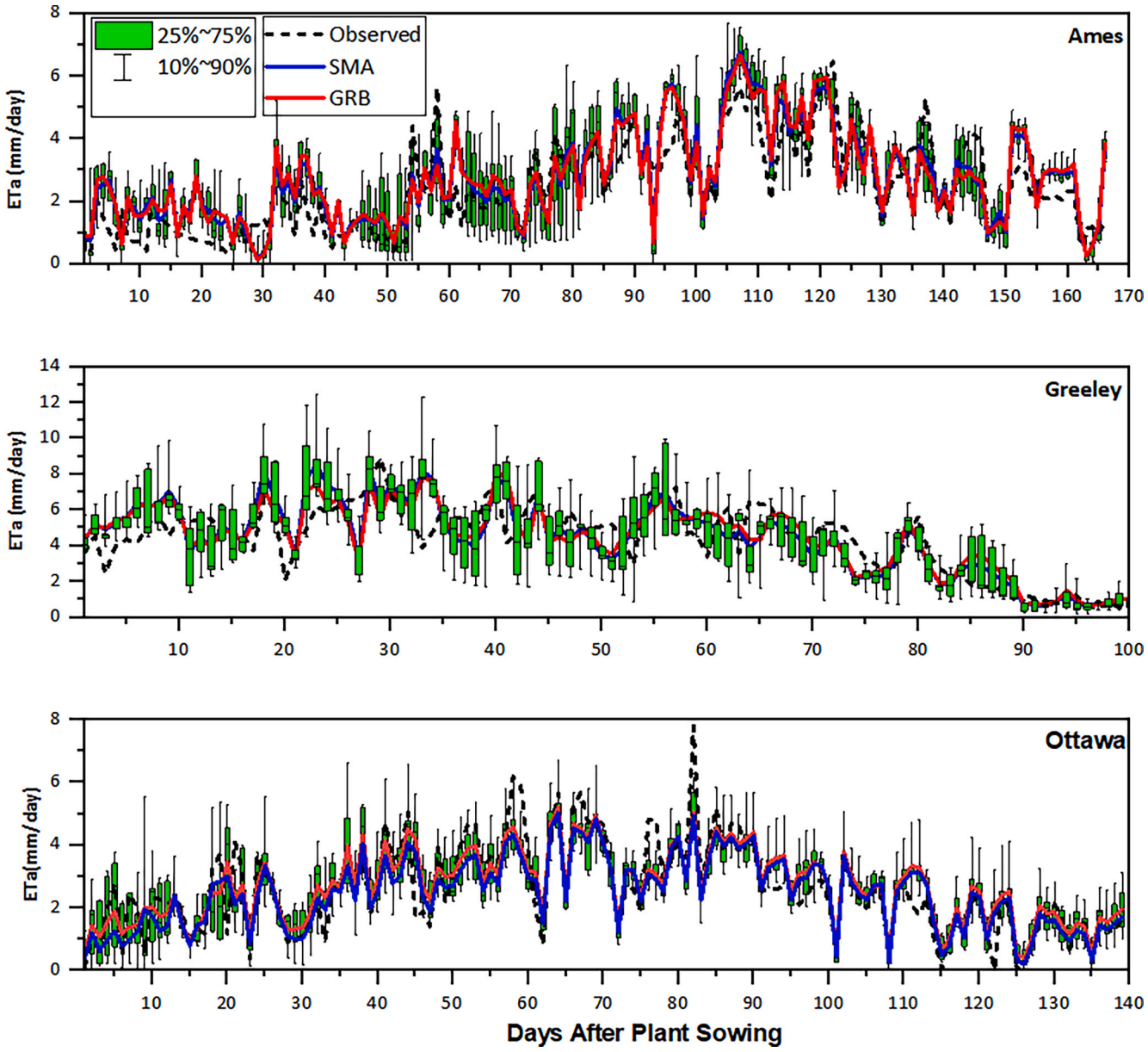



**Fig. 5.** Box plots of daily simulated evapotranspiration (ETa) of the maize season 2006, 2010, and 2006 across five maize models at Group A sites (Ames, Greeley, and Ottawa) respectively. Observed daily ETa values, and the GRB and SMA multi-model averaging approach derived daily ETa values from the five maize models are also presented. The simulated outputs of the calibrated phase where all maize models were fully calibrated using crop phenology dates, LAI, soil moisture, ETa and yield data.

substantially when simulated yields were averaged using model-averaging methods at all sites. The GRA performed the best at all sites, followed by GRC, GRB, IR, BGA, SMA, and the median. The RRMSE between ensemble and measured yields ranged from 0.03 to 4.0 % at Mead irrigated, 5.6 to 12.8 % at Mead rainfed, 4.2 to 15 % at Bushland 100 % MESA, and 2.8 to 19 % at Bushland 75 % MESA sites across all model-averaging methods (Table 4). Additionally, the ensembling of simulated yield from group maize models showed mixed results compared to combining simulated yields from all maize models across all model-averaging methods. There was a marginal improvement in yield simulation at Mead rainfed and Bushland 75 % MESA sites compared to all maize models, while there was a slight decrease noted at Mead irrigated and Bushland 100 % MESA sites (Table 4).

## 4. Discussion

### 4.1. Blind vs calibrated

Combining simulations from multiple models through various model-averaging approaches often provides more accurate simulation performance (Sándor et al., 2023). In this study, as anticipated, MAAs performed slightly better during the calibrated phase than for the blind phase for combining daily ETa and yield simulations of all and group maize models (Tables 5 and 6). In crop modeling, calibration is a crucial process aimed at estimating unknown parameters using field observations, thereby reducing uncertainty in model simulations and making predictions more reliable (He et al., 2017). MAAs tend to perform better in the calibrated phase because the models are fine-tuned to specific datasets, which minimizes errors and variance, resulting in more accurate and stable predictions (Fletcher, 2018).

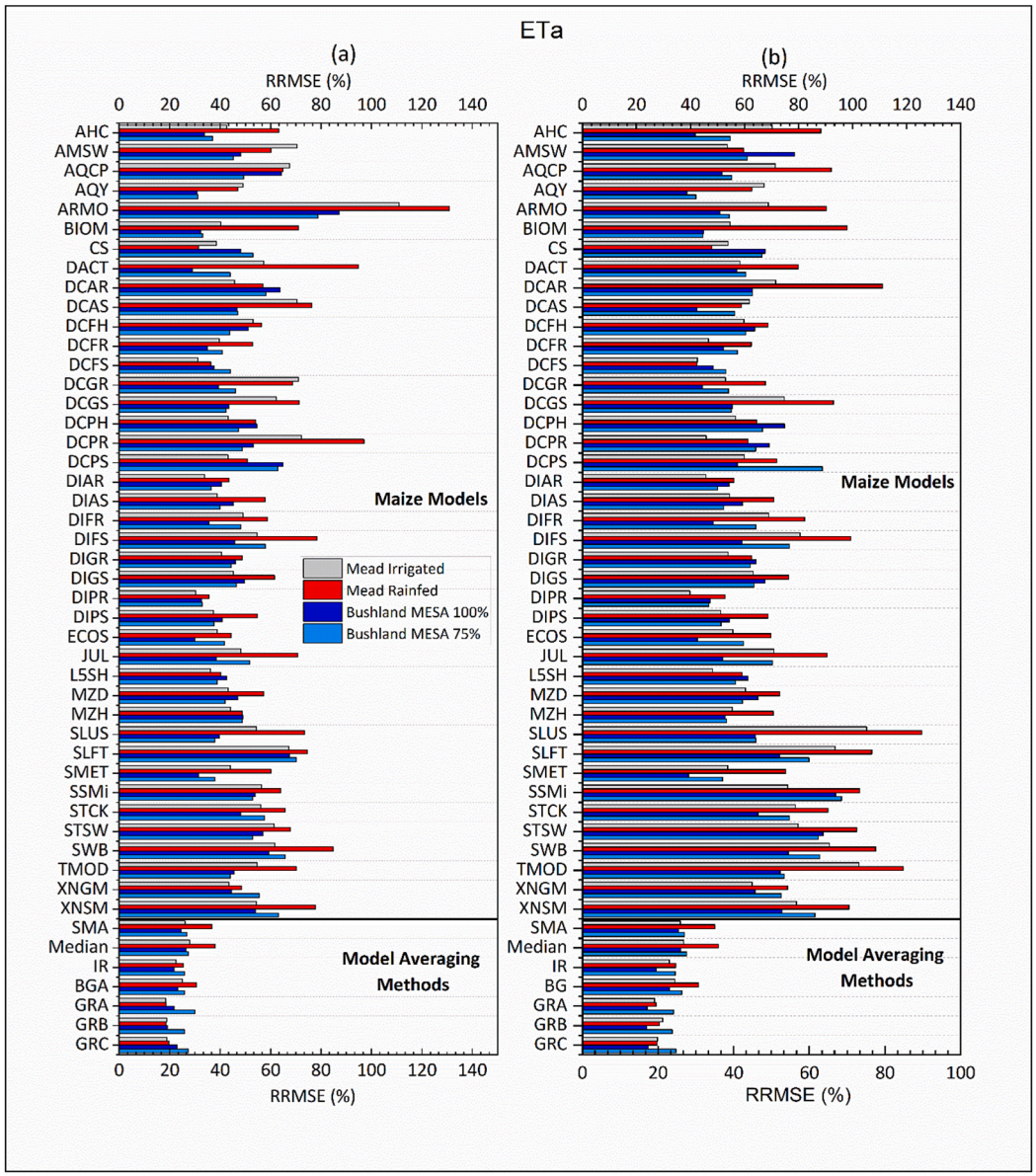



**Fig. 6.** RRMSE between the measured and simulated daily crop evapotranspiration (ETa) across 41 maize models and seven multi-model averaging approaches (MMA) at Group B sites under uncalibrated (a) and calibration phase (b).

Interestingly, MAAs also performed well in the blind phase. The outcomes of the present study are comparable to those of Bassu et al. (2014) and Kimball et al. (2019), where the maize yield and ETa simulations from uncalibrated maize models in different climatic conditions sites were combined using the mean and median. However, in this study, an additional five MAAs were tested, which will be discussed in the next section. Similarly, Ajami et al. (2006) found that averaging streamflow simulations of uncalibrated multiple hydrological models using four model combination methods performed better than a calibrated single hydrological model. These studies found that multi-model combinations could enhance prediction accuracy by compensating for individual model errors to reduce variance (Bassu et al., 2014; Kimball et al., 2019; Kimball et al., 2023; Sándor et al., 2023; Couëdel et al., 2024). The multi-model combination improves the simulation accuracy by reducing the variance associated with the predictions (Bassu et al., 2014; Fletcher, 2018). The individual model might exhibit high variance due to their sensitivity to model structures and parameters. By averaging the outputs of multiple models, these variances are reduced, leading to more stable and reliable predictions. In addition, different models may make different errors when predicting. When these models are averaged, the errors can cancel each other out to some extent, resulting in a more accurate overall prediction. Nonetheless, while multi-model ensembles offer a way to learn from the errors across various studies and improve the models, some individual models might still outperform the mean and median (Kothari et al., 2022).

### 4.2. *Best model averaging method for ETa and yield*

The study assesses how well different MAA can reduce variability and improve the accuracy of daily ETa and yield simulations at Group A and Group B sites. Remarkably, SMA and the median approach performed better than individual calibrated maize models in 98 % of the cases during the blind phase at Group A sites, with SMA usually outperforming the median. Similar results were observed in Group B sites for ETa and yield. This could be due to a trade-off in prediction errors among different models, leading to more accurate overall predictions. These findings are comparable to those of Ajami et al. (2006), Bassu et al. (2014), Arsenault et al. (2015), Sándor et al. (2023), and Couëdel

**Table 4**
Comparison of RRMSE between (a) measured daily ETa and ensembled daily ETa, and (b) measured maize yield and ensembled maize yield for all maize models and group maize models using seven multi-model averaging approaches (MAA) at Group B sites under Blind and Calibrated Phases.

| Averaging approaches | Blind | | | | | | | Calibrated | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | All models | | | | Group models | | | All models | | | | Group models | | |
| | Mead Irrigated | Mead Rainfed | Bushland 100 % MESA | Bushland 75 % MESA | Mead Irrigated | Mead Rainfed | Bushland 100 % MESA | Bushland 75 % MESA | Mead Irrigated | Mead Rainfed | Bushland 100 % MESA | Bushland 75 % MESA | Mead Irrigated | Mead Rainfed | Bushland 100 % MESA | Bushland 75 % MESA |
| **Daily ETa (a)** | | | | | | | | | | | | | | | | |
| SMA | 26.1 | 36.8 | 24.6 | 26.8 | 24.4 | 34.3 | 23.5 | 25.7 | 25.9 | 35.0 | 25.4 | 26.9 | 23.8 | 33.5 | 24.8 | 26.1 |
| Median | 28.0 | 38.1 | 26.4 | 27.5 | 24.4 | 34.6 | 25.7 | 27.2 | 26.7 | 35.9 | 26.0 | 27.6 | 23.6 | 33.8 | 25.6 | 27.1 |
| IR | 22.5 | 25.4 | 21.7 | 25.8 | 21.2 | 24.0 | 20.3 | 24.9 | 22.9 | 24.7 | 19.4 | 24.5 | 20.8 | 23.9 | 18.9 | 24.4 |
| BGA | 25.0 | 30.6 | 23.3 | 25.9 | 22.7 | 27.4 | 22.1 | 25.4 | 24.4 | 30.7 | 23.0 | 26.2 | 22.4 | 28.6 | 22.7 | 25.9 |
| GRA | 18.4 | 18.7 | 21.8 | 30.0 | 18.6 | 18.1 | 19.5 | 30.3 | 19.0 | 19.4 | 17.1 | 24.1 | 19.4 | 19.4 | 17.0 | 24.0 |
| GRB | 18.9 | 18.5 | 19.0 | 25.9 | 18.9 | 18.5 | 16.9 | 22.2 | 21.2 | 20.4 | 17.0 | 23.7 | 19.7 | 20.5 | 16.2 | 21.5 |
| GRC | 18.9 | 19.7 | 22.9 | 27.4 | 18.8 | 18.3 | 21.6 | 27.9 | 19.8 | 19.7 | 17.4 | 24.7 | 19.9 | 20.0 | 17.4 | 25.0 |
| **Seasonal yield (b)** | | | | | | | | | | | | | | | | |
| SMA | 8.9 | 14.0 | 26.0 | 9.6 | 7.4 | 11.7 | 28.3 | 11.8 | 4.0 | 12.4 | 15.0 | 15.8 | 5.1 | 9.8 | 14.9 | 16.6 |
| Median | 13.3 | 17.0 | 20.3 | 11.0 | 10.0 | 16.0 | 26.3 | 11.9 | 1.6 | 12.8 | 12.8 | 19.1 | 2.0 | 9.3 | 12.7 | 16.8 |
| IR | 2.6 | 6.5 | 10.7 | 2.8 | 3.8 | 6.5 | 10.1 | 16.8 | 0.2 | 6.2 | 7.3 | 4.2 | 0.4 | 7.3 | 7.2 | 4.2 |
| BGA | 2.0 | 6.4 | 11.0 | 2.8 | 2.7 | 6.4 | 10.2 | 15.8 | 0.1 | 6.2 | 7.7 | 4.1 | 0.3 | 6.8 | 7.5 | 4.1 |
| GRA | 7.9 | 1.6 | 6.8 | 1.7 | 2.0 | 2.4 | 8.0 | 8.3 | 0.0 | 5.6 | 4.2 | 2.8 | 0.1 | 3.6 | 3.5 | 3.4 |
| GRB | 1.5 | 4.7 | 10.7 | 2.8 | 2.7 | 4.7 | 9.1 | 15.5 | 0.1 | 5.7 | 6.6 | 4.2 | 0.2 | 5.9 | 6.6 | 4.2 |
| GRC | 8.4 | 1.9 | 7.6 | 1.9 | 2.4 | 4.2 | 8.3 | 10.1 | 0.1 | 7.2 | 4.8 | 4.1 | 0.3 | 5.9 | 5.9 | 4.1 |

et al. (2024), which showed that the mean of simulated streamflow and yield from hydrological and crop models, respectively, was better than individual calibrated models.

Further enhancement in daily ETa and maize yield simulations was noted when other model averaging methods, such as IR, BGA, GRA, GRB, and GRC, were used. Overall, the improvements ranged between 3.5 and 6.5 % for daily ETa and 3.3–9.7 % in terms of RRMSE for yield simulations at Group A sites across the five MAAs compared to the median (Table 5). Similarly, improvements in daily ETa and yield simulations ranged between 3.2 % and 8.7 %, and 7.3 % and 9.5 %, respectively, at Group B sites (Table 6).). The improvement in daily ETa and yield estimations by the additional five MAAs over the median was slightly greater for daily ETa and moderately greater for yield in the blind phase compared to the calibrated phase (Table 5 and Table 6). BGA often performed better in combining daily ETa simulations than SMA and the median, though it was usually outperformed by its variant IR (Tables 5 and 6). This can be explained by the IR method's disregard for outliers (Aiolfi and Timmermann, 2006). For yield simulations, BGA and IR showed almost similar performance. According to Diks and Vrugt (2010), BGA did not outperform other methods (AICA, BICA, BMA, and MLR A) except SMA.

When comparing the performance of GRA, GRB, and GRC, there were only marginal differences in their ability to combine daily ETa and yield simulations in 75 % of cases, aligning with the study by Arsenault et al. (2015) (Tables 2 and 4). GRA, GRB, and GRC performed considerably better than SMA and the median and slightly to moderately better than IR and BGA, depending on the site. Overall, averaging the RRMSE of all sites for all maize models and group maize models for blind and calibrated phases revealed that GRB was best for ensemble of daily ETa simulations, while GRA was best for yield simulations (Tables 5 and 6). GRB slightly outperformed GRA by 0.5–1.5 % in terms of RRMSE to ensemble daily ETa, depending on the site and model group. GRA clearly outperformed GRB by a 2–3 % in terms of RRMSE, which is notable given the lower error ranges typically associated with yield. GRB improved daily ETa estimation by an average of 4 % and 8.5 % in terms of RRMSE than the median, while GRA enhanced maize yield estimation by 7.5 % and 10.9 % for Group A and Group B sites, respectively. The consistent performance of GRA and GRB in ensembling yield and daily ETa across different soil types, climatic conditions, crop management practices and model ensembles supports the strength of our findings within the context of maize crop, however these findings may not generalize to other crops (e.g., soybean, wheat, canola, potato, alfalfa) or regions, and it need to be examined in future studies.

This is likely because of higher bias in daily ETa simulations across maize models compared to yield simulations. GRA was better at reducing variance in yield simulations due to incorporating variance reduction. In contrast, GRB reduces variance by giving positive higher weights to well-performing models while minimum weight to the worst-performing models even in some cases zero. Therefore, it combined the daily ETa simulations slightly better than other MAA. For ETa, the results were contradicted by Ajami et al. (2006), Arsenault et al. (2015), and Wan et al. (2021) and were comparable to Kumar et al. (2015).

Kumar et al. (2015) found that GRB was the best method for combining simulated river discharge from eight hydrological models. For crop yield, findings were in line with (Diks and Vrugt, 2010), who reported that GRA's results were similar to advanced MAA such as Bayesian Model Averaging (BMA) and Mallows Model Averaging (MAAS). The advantage of using GRA over BMA or MAAS can be notable since GRA has straightforward solutions for determining weights. In contrast, finding the best weights for BMA and MAAS requires more complex and time-consuming methods, such as the Differential Evolution Adaptive Metropolis (DREAM) and adaptive Markov chain Monte Carlo (MCMC) algorithm.

Overall, the GRA and GRB methods were found to outperform others for ensemble yield and ETa simulations of maize models, respectively, in both data sets. This emphasizes the importance of selecting appropriate

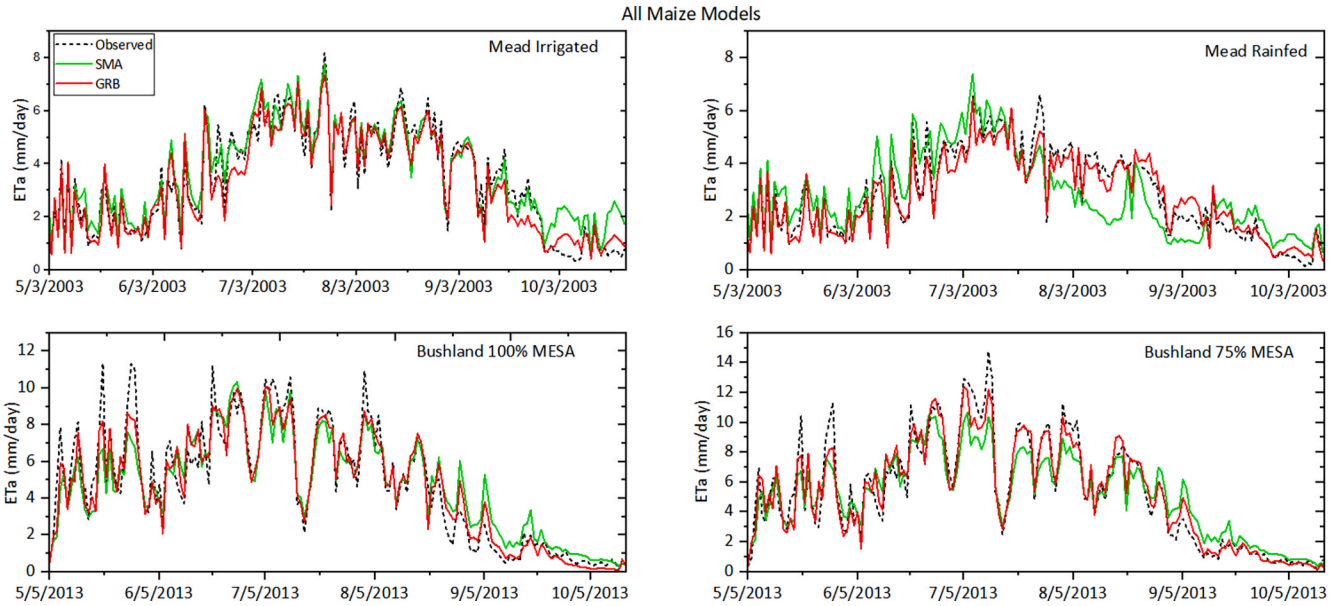



**Fig. 7.** A comparison of measured daily ETa and an ensemble of daily ETa simulations of all maize models using SMA and GRB multi-model averaging approaches (MMA) at Group B sites under uncalibrated phase.

averaging techniques. The success of these methods can be attributed to their ability to integrate multiple model outputs, leveraging the strengths and compensating for the weaknesses of individual models.

Moreover, ensemble group maize models improved the simulation accuracy of crop yield and ETa in a few cases compared to ensemble all maize models. However, the accuracy of the ensembled ETa and yield simulation of group maize models was similar to that of the ensembled ETa and yield simulation of all maize models. This finding suggests that the diversity of models in the ensemble plays a crucial role in enhancing prediction accuracy. Therefore, it is advisable to select ensemble members from different crop family models to achieve the best results, although its also true that the quality of modelers regarding the assumptions they make in parameterizing models is also of importance (Albanito et al., 2022).

### *4.3. Model averaging methods when "no observations data" is available*

Most MAA, such as IR, BGA, GRA, GRB, and GRC, typically rely on ground measurement data to determine the weights for each model in the ensemble. This data is crucial for selecting the best models and assigning appropriate weights. However, in real-world scenarios, experimental data not be available, posing substantial challenges for model selection and weighting.

In such situations, SMA and the median method have shown promising results. SMA and the median method are straightforward approaches that average predictions from multiple models by assigning equal weights to each. This simplicity is particularly advantageous when there is no prior information about the performance of the individual models. By averaging the outputs, SMA reduces the impact of biases or errors from any single model, leading to more robust overall predictions. Both methods were effective in the current study, where they combined multiple crop model outputs to improve predictions of daily ETa and yield, even in the blind phase. This finding is consistent with previous crop modeling studies by Bassu et al. (2014), Martre et al. (2015), Kothari et al. (2022), Kimball et al. (2019, 2023), who reported that the mean and median of ETa and yield simulations from multiple crop models often outperform individual crop models.

However, the main drawback of SMA and the median method is that they do not fully leverage the strengths of the better-performing models. Because all models are weighted equally, these methods may underutilize the models that have superior predictive capabilities. Despite this limitation, SMA and the median method remain valuable tools in scenarios where observational data are lacking, providing a practical means of improving predictive accuracy by mitigating individual model weaknesses.

## 5. Conclusions

Averaging the results from multiple agricultural systems models has shown high accuracy in predicting crop yield and ETa. However, among those available Model Averaging Approaches (MAA), it is not known which one performed the best. Therefore, this study aimed to evaluate the performance of seven MAA (SMA, Median, IR, BGA, GRA, GRB, and GRC) across eleven sites in North America to predict maize yield and daily ETa using two ensemble-size maize crop models (all maize models and group maize models) and two calibration approaches (Blind and Calibrated phases). The data come from two sources: simulations for Group A sites were done in this study, while simulations for Group B sites were carried out by the Maize AgMIP project team.

The following conclusions were drawn from the study:

- **Model Averaging Approaches:** All MAA (Model Averaging Approaches) generally performed well, often surpassing individual crop models during both the blind and calibration phases. Among the MAA, the GRB method typically provided the closest match to measured daily ETa values, while the GRA method was most accurate for maize yield across all sites and phases. The simple mean consistently outperformed the median at all sites. Therefore, GRA and GRB are recommended for averaging simulations of yield and ETa, respectively, when measured data is available. However, in the absence of observed ETa and yield data, the SMA method can be used to ensemble the yield and ETa simulations.
- **Individual maize model performance:** No single maize model consistently performed best at all sites for simulating yield and daily ETa. Results indicate that fully calibrating the crop model, slightly to significantly improved the daily ETa and yields simulations compared to the blind phase, depending on maize models and sites.
- **Phase comparison for modeling averaging:** The performance of all MAA improved slightly to moderately for daily ETa and yield from the blind phase to the calibrated phase across all sites.

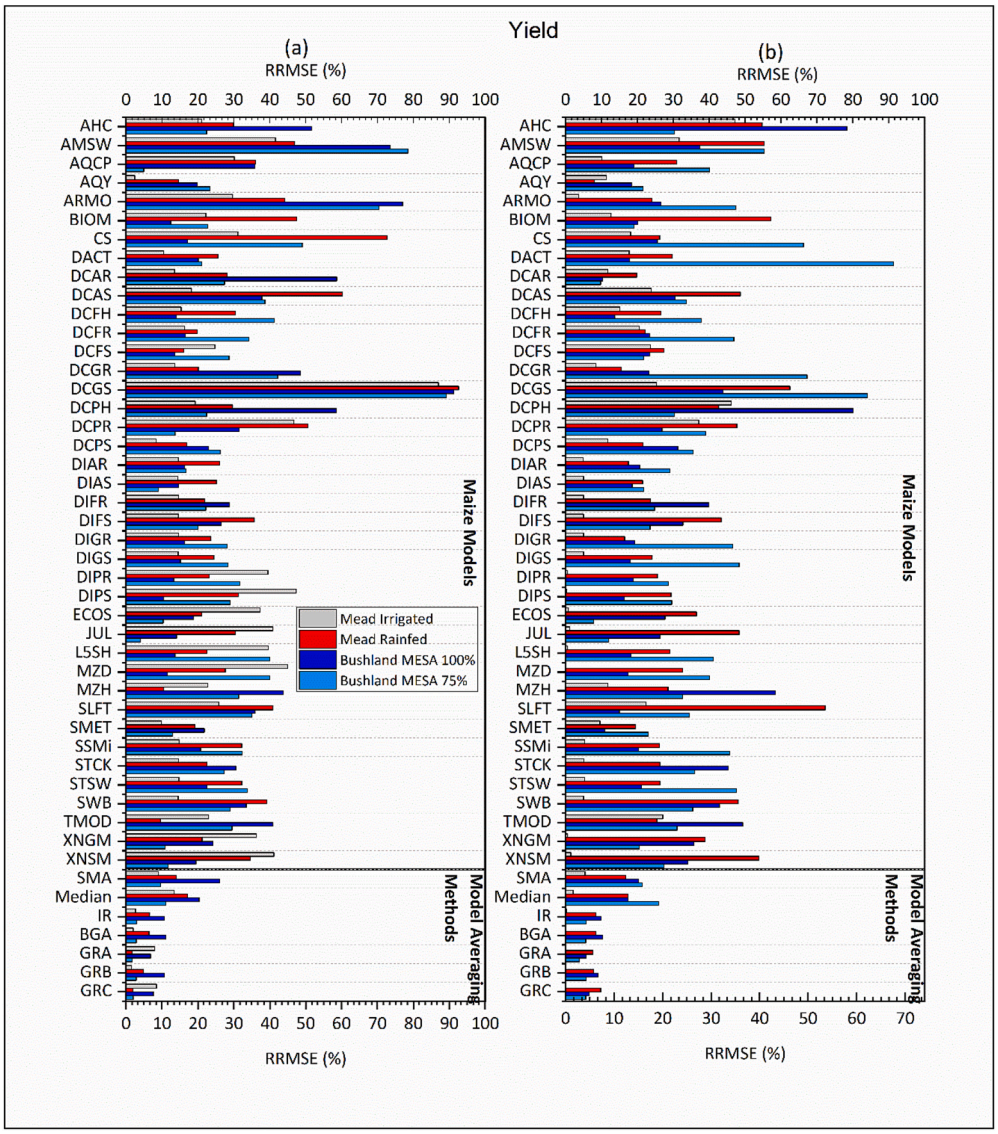



**Fig. 8.** RRMSE between the measured and simulated maize yield across maize models and multi-model averaging approaches (MMA) at Group B sites under uncalibrated (a) and calibration phase (b).

- **Ensemble member models**: Using an ensemble of group maize models with different model structures slightly enhanced the accuracy of daily ETa and yield simulations at Group B in comparison to using an ensemble of all maize models.

These findings highlight the potential of MAA to improve the precision of maize yield and daily ETa estimates, emphasizing the importance of using diverse model ensembles to achieve accurate agricultural predictions. However, these findings may be limited to maize crop in North America. The applicability of these MAA methods to other crops (e.g., soybean, wheat, canola, potato, alfalfa) or regions still need to be examined, as their performance may differ.

## CRediT authorship contribution statement

**Viveka Nand:** Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Conceptualization. **Zhiming Qi:** Writing – review & editing, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **Liwang Ma:** Writing – review & editing, Data curation. **Matthew J. Helmers:** Data curation. **Chandra A. Madramootoo:** Data curation. **Ward N. Smith:** Writing – review & editing, Data curation. **Tiequan Zhang:** Data curation. **Tobias K.D. Weber:** Writing – review & editing, Supervision, Software. **Elizabeth Pattey:** Data curation. **Ziwei Li:** Software. **Jiaxin Wang:** Software. **Virginia L. Jin:** Data curation. **Qianjing Jiang:** Software. **Mario Tenuta:** Data curation. **Thomas J. Trout:** Data curation. **Haomiao Cheng:** Software. **R. Daren Harmel:** Writing – review & editing, Software. **Bruce A. Kimball:** Writing – review & editing, Software, Data curation. **Kelly R. Thorp:** Writing – review & editing, Software. **Kenneth J. Boote:** Software. **Claudio Stockle:** Software. **Andrew E. Suyker:** Data curation. **Steven R. Evett:** Data curation. **David K. Brauer:** Data curation. **Gwen G. Coyle:** Data curation. **Karen S. Copeland:** Data curation. **Gary W. Marek:** Data curation. **Paul D. Colaizzi:** Data curation. **Marco Acutis:** Software. **Seyyed Majid Alimagham:** Software. **Sotirios Archontoulis:** Software. **Faye Babacar:** Software. **Zoltán Barcza:** Writing – review & editing, Software. **Bruno Basso:** Software. **Patrick Bertuzzi:** Software. **Julie Constantin:** Software. **Massimiliano De Antoni Migliorati:** Software. **Benjamin Dumont:** Software. **Jean-Louis Durand:** Software. **Nándor Fodor:** Software. **Thomas Gaiser:** Software. **Pasquale Garofalo:** Software. **Sebastian Gayler:** Software. **Luisa Giglio:** Software. **Robert Grant:** Software. **Kaiyu Guan:** Software. **Gerrit Hoogenboom:**

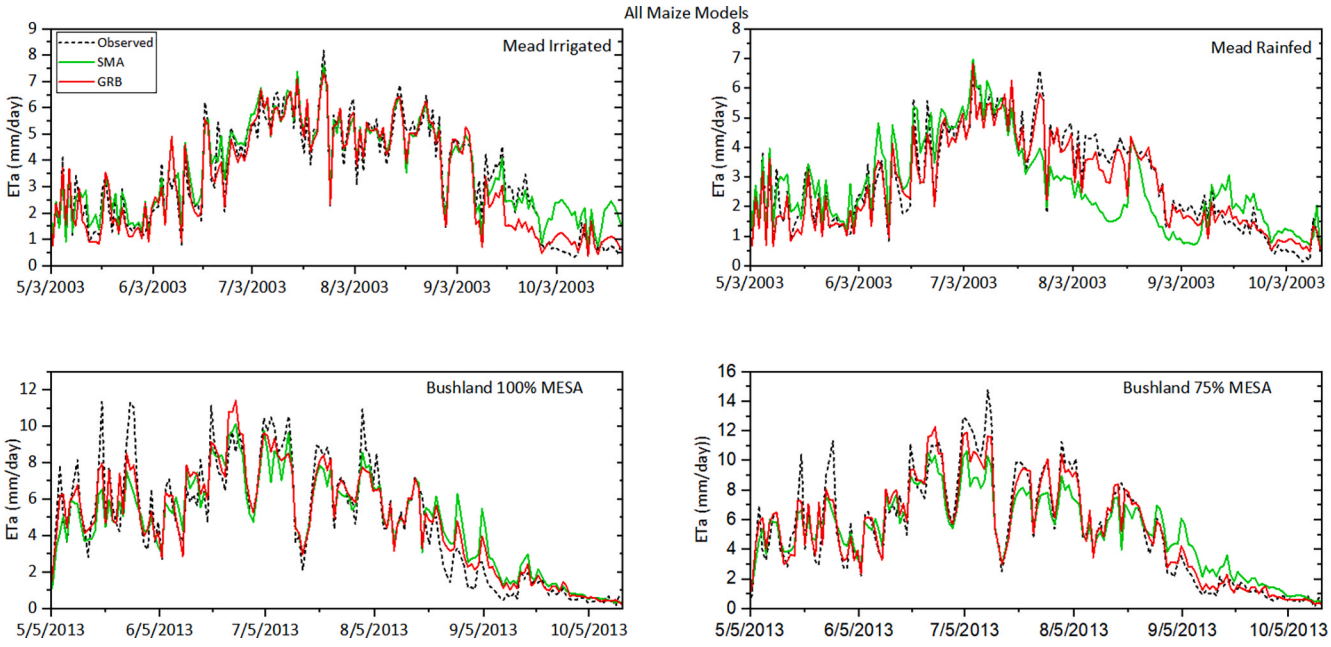



**Fig. 9.** A comparison of measured daily ETa simulations and an ensemble of daily ETa simulations of all maize models using SMA and GRB multi-model averaging approaches (MMA) at Group B sites under calibration phase.

**Table 5**
Average RRMSE between measured daily ETa and maize yield, and ensembled daily ETa and maize yield, respectively, for all maize models and group maize models using multi-model averaging approaches (MAA) at Group A sites for both the blind and calibration phases.

| | Daily ETa | | | | | Seasonal Yield | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Blind | | Calibrated | | Overall | Blind | | Calibrated | | Overall |
| | All models | Group models | All models | Group models | | All models | Group models | All models | Group models | |
| SMA | 39.5 | 37.6 | 33.0 | 33.5 | 35.9 | 17.3 | 18.3 | 8.8 | 9.2 | 13.4 |
| Median | 42.6 | 41.1 | 33.9 | 34.6 | 38.1 | 17.8 | 19.6 | 8.4 | 10.0 | 14.0 |
| IR | 37.0 | 37.2 | 32.8 | 33.5 | 35.1 | 13.1 | 14.1 | 6.9 | 8.5 | 10.7 |
| BGA | 36.6 | 36.8 | 32.4 | 33.3 | 34.8 | 12.3 | 12.6 | 6.7 | 8.0 | 9.9 |
| GRA | 34.5 | 34.7 | 32.7 | 33.6 | 33.9 | 3.4 | 4.9 | 3.2 | 4.0 | 3.9 |
| GRB | 35.4 | 36.6 | 31.9 | 33.1 | 34.2 | 10.8 | 12.0 | 5.5 | 7.8 | 9.0 |
| GRC | 33.3 | 33.9 | 33.3 | 33.5 | 33.5 | 4.0 | 5.6 | 4.6 | 5.5 | 4.9 |
| Mean | 37.0 | 36.9 | 32.9 | 33.6 | 35.1 | 11.2 | 12.5 | 6.3 | 7.6 | 9.4 |

**Table 6**
Average RRMSE between measured daily ETa and yield, and ensembled daily ETa and yield, respectively, for all maize models and group maize models using multi-model averaging approaches (MAA) at Group B sites for both the blind and calibration phases.

| Averaging approaches | Daily ETa | | | | | Yield | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Blind | | Calibrated | | Overall | Blind | | Calibrated | | Overall |
| | All models | Group models | All models | Group models | | All models | Group models | All models | Group models | |
| SMA | 28.6 | 27.0 | 28.3 | 27.0 | 27.7 | 14.6 | 14.8 | 11.8 | 11.6 | 13.2 |
| Median | 30.0 | 27.9 | 29.0 | 27.5 | 28.6 | 15.4 | 16.1 | 11.6 | 10.2 | 13.3 |
| IR | 23.9 | 22.6 | 22.9 | 22.0 | 22.8 | 5.6 | 9.3 | 4.5 | 4.8 | 6.0 |
| BGA | 26.2 | 24.4 | 26.1 | 24.9 | 25.4 | 5.5 | 8.8 | 4.5 | 4.7 | 5.9 |
| GRA | 22.2 | 21.6 | 19.9 | 20.0 | 20.9 | 4.5 | 5.2 | 3.2 | 2.7 | 3.9 |
| GRB | 20.6 | 19.1 | 20.6 | 19.5 | 19.9 | 4.9 | 8.0 | 4.1 | 4.2 | 5.3 |
| GRC | 22.2 | 21.7 | 20.4 | 20.6 | 21.2 | 4.9 | 6.3 | 4.1 | 4.0 | 4.8 |
| Mean | 24.8 | 23.5 | 23.9 | 23.1 | 23.8 | 7.9 | 9.8 | 6.2 | 6.0 | 7.5 |

Software. **Soo-Hyung Kim:** Software. **Isaya Kisekka:** Software. **Jon Lizaso:** Software. **Sara Masia:** Software. **Huimin Meng:** Software. **Valentina Mereu:** Software. **Ahmed Mukhtar:** Software. **Alessia Perego:** Software. **Bin Peng:** Software. **Eckart Priesack:** Software. **Vakhtang Shelia:** Software. **Richard Snyder:** Software. **Afshin Soltani:** Software. **Donatella Spano:** Software. **Amit Srivastava:** Software. **Aimee Thomson:** Software. **Dennis Timlin:** Writing – review & editing, Software. **Antonio Trabucco:** Software. **Heidi Webber:** Software. **Magali Willaume:** Software. **Karina Williams:** Software. **Michael van der Laan:** Software. **Domenico Ventrella:** Software. **Michelle Viswanathan:** Software. **Xu Xu:** Software. **Wang Zhou:** Software.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

We are grateful to the Ministry of Social Justice and Empowerment, Government of India (11015/48/2018-SCD-V), McGill University, and the Natural Sciences and Engineering Research Council of Canada (NSERC) for providing financial support for the first author to carry out this study.

## Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jhydrol.2025.133631.

## Data availability

Data will be made available on request.

## References

Aiolfi, M., Timmermann, A., 2006. Persistence in forecasting performance and conditional combination strategies. J. Econom. 135 (1), 31–53. https://doi.org/10.1016/j.jeconom.2005.07.015.

Ajami, N.K., Duan, Q., Gao, X., Sorooshian, S., 2006. Multimodel combination techniques for analysis of hydrological simulations: application to distributed model intercomparison project results. J. Hydrometeorol. 7 (4), 755–768. https://doi.org/10.1175/JHM519.1.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19 (6), 716–723. https://doi.org/10.1109/TAC.1974.1100705.

Albanito, F., McBey, D., Harrison, M., Smith, P., Ehrhardt, F., Bhatia, A., Bellocchi, G., Brilli, L., Carozzi, M., Christie, K., Doltra, J., 2022. How modelers model: the overlooked social and human dimensions in model intercomparison studies. Environ. Sci. Technol. 56 (18), 13485–13498. https://doi.org/10.1021/acs.est.2c02023.

Arsenault, R., Gatien, P., Renaud, B., Brissette, F., Martel, J., 2015. A comparative analysis of 9 multi-model averaging approaches in hydrological continuous streamflow simulation. J. Hydrol. 529, 754–767. https://doi.org/10.1016/j.jhydrol.2015.09.001.

Bassu, S., Brisson, N., Durand, J., Boote, K., Lizaso, J., Jones, J.W., Rosenzweig, C., Ruane, A.C., Adam, M., Baron, C., Basso, B., Biernath, C., Boogaard, H., Conijn, S., Corbeels, M., Deryng, D., De Sanctis, G., Gayler, S., Grassini, P., Hatfield, J., Hoek, S., Izaurralde, C., Jongschaap, R., Kemanian, A.R., Kersebaum, K.C., Kim, S., Kumar, N. S., Makowski, D., Müller, C., Nendel, C., Priesack, E., Pravia, M.V., Sau, F., Shcherbak, I., Tao, F., Teixeira, E., Timlin, D., Waha, K., 2014. How do various maize crop models vary in their responses to climate change factors? Glob. Change Biol. 20 (7), 2301–2320. https://doi.org/10.1111/gcb.12520.

Cheng, H., Shu, K., Qi, Z., Ma, L., Jin, V.L., Li, Y., Schmer, M.R., Wienhold, B.J., Feng, S., 2021. Effects of residue removal and tillage on greenhouse gas emissions in continuous corn systems as simulated with RZWQM2. J. Environ. Manag. 285, 112097. https://doi.org/10.1016/j.jenvman.2021.112097.

Couëdel, A., Falconnier, G.N., Adam, M., Cardinael, R., Boote, K., Justes, E., Smith, W.N., Whitbread, A.M., Affholder, F., Balkovic, J., Basso, B., 2024. Long-term soil organic carbon and crop yield feedbacks differ between 16 soil-crop models in sub-Saharan Africa. Eur. J. Agron. 155, 127109. https://doi.org/10.1016/j.eja.2024.127109.

Crépeau, M., Jégo, G., Morissette, R., Pattey, E., Morrison, M.J., 2021. Predictions of soybean harvest index evolution and evapotranspiration using STICS crop model. Agron. J. 113 (4), 3281–3298. https://doi.org/10.1002/agj2.20765.

Deb, P., Moradkhani, H., Han, X., Abbaszadeh, P., Xu, L., 2022. Assessing irrigation mitigating drought impacts on crop yields with an integrated modeling framework. J. Hydrol. 609, 127760. https://doi.org/10.1016/j.jhydrol.2022.127760.

Diks, C.G., Vrugt, J.A., 2010. Comparison of point forecast accuracy of model averaging methods in hydrologic applications. Stoch. Environ. Res. Risk Assess. 24, 809–820. https://doi.org/10.1007/s00477-010-0378-z.

Fang, Q., Ma, L., Harmel, R.D., Yu, Q., Sima, M.W., Bartling, P.N.S., Malone, R.W., Nolan, B.T., Doherty, J., 2019. Uncertainty of CERES-Maize calibration under different irrigation strategies using PEST optimization algorithm. Agronomy 9 (241), 1–17. https://www.mdpi.com/2073-4395/9/5/241.

Fletcher, D., 2018. In: Why Model Averaging?. Springer, Berlin Heidelberg, pp. 1–29.

Gao, Y., Wallach, D., Hasegawa, T., Tang, L., Zhang, R., Asseng, S., Kahveci, T., Liu, L., He, J., Hoogenboom, G., 2021. Evaluation of crop model prediction and uncertainty using Bayesian parameter estimation and Bayesian model averaging. Agric. For. Meteorol. 311, 108686. https://doi.org/10.1016/j.agrformet.2021.108686.

Granger, C.W.J., Ramanathan, R., 1984. Improved methods of combining forecasts. J. Forecast. 3 (2), 197–204. https://doi.org/10.1002/for.3980030207.

He, D., Wang, E., Wang, J., Robertson, M.J., 2017. Data requirement for effective calibration of process-based crop models. Agric. For. Meteorol. 234, 136–148. https://doi.org/10.1016/j.agrformet.2016.12.015.

Huang, X., Huang, G., Yu, C., Ni, S., Yu, L., 2017. A multiple crop model ensemble for improving broad-scale yield prediction using Bayesian model averaging. Field Crops Res. 211, 114–124. https://doi.org/10.1016/j.fcr.2017.06.011.

Jafarzadeh, A., Khashei-Siuki, A., Pourreza-Bilondi, M., 2022. Performance assessment of model averaging techniques to reduce structural uncertainty of groundwater modeling. Water Resour. Manag. 36 (1), 353–377. https://doi.org/10.1007/s11269-021-03031-x.

Jamieson, P.D., Porter, J.R., Wilson, D.R., 1991. A test of the computer simulation model ARCWHEAT on wheat crops grown in New Zealand. Field Crops Res. 27 (4), 337–350. https://doi.org/10.1016/0378-4290(91)90040-3.

Jiang, Q., Madramootoo, C.A., Qi, Z., 2022. Soil carbon and nitrous oxide dynamics in corn (Zea mays L.) production under different nitrogen, tillage and residue management practices. Field Crops Res. 277, 108421.

Jiang, Q., Qi, Z., Lu, C., Tan, C.S., Zhang, T., Prasher, S.O., 2020. Evaluating RZ-SHAW model for simulating surface runoff and subsurface tile drainage under regular and controlled drainage with subirrigation in southern Ontario. Agric. Water Manag. 237, 106179. https://doi.org/10.1016/j.agwat.2020.106179.

Kimball, B.A., Boote, K.J., Hatfield, J.L., Ahuja, L.R., Stockle, C., Archontoulis, S., Baron, C., Basso, B., Bertuzzi, P., Constantin, J., Deryng, D., Dumont, B., Durand, J.-L., Ewert, F., Gaiser, T., Gayler, S., Hoffmann, M.P., Jiang, Q., Kim, S.-H., Lizaso, J., Moulin, S., Nendel, C., Parker, P., Palosuo, T., Priesack, E., Qi, Z., Srivastava, A., Stella, T., Tao, F., Thorp, K.R., Timlin, D., Twine, T.E., Webber, H., Willaume, M., Williams, K., 2019. Simulation of maize evapotranspiration: an inter-comparison among 29 maize models. Agric. For. Meteorol. 271, 264–284. https://doi.org/10.1016/j.agrformet.2019.02.037.

Kimball, B.A., Thorp, K.R., Boote, K.J., Stockle, C., Suyker, A.E., Evett, S.R., Brauer, D.K., Coyle, G.G., Copeland, K.S., Marek, G.W., Colaizzi, P.D., Acutis, M., Alimagham, S., Archontoulis, S., Babacar, F., Barcza, Z., Basso, B., Bertuzzi, P., Constantin, J., De Antoni Migliorati, M., Dumont, B., Durand, J., Fodor, N., Gaiser, T., Garofalo, P., Gayler, S., Giglio, L., Grant, R., Guan, K., Hoogenboom, G., Jiang, Q., Kim, S., Kisekka, I., Lizaso, J., Masia, S., Meng, H., Mereu, V., Mukhtar, A., Perego, A., Peng, B., Priesack, E., Qi, Z., Shelia, V., Snyder, R., Soltani, A., Spano, D., Srivastava, A., Thomson, A., Timlin, D.J., Trabucco, A., Webber, H., Weber, T., Willaume, M., Williams, K., van der Laan, M., Ventrella, D., Viswanathan, M., Xu, X., Zhou, W., 2023. Simulation of evapotranspiration and yield of maize: an inter-comparison among 41 maize models. Agric. For. Meteorol. 333, 109396. https://doi.org/10.1016/j.agrformet.2023.109396.

Kothari, K., Battisti, R., Boote, K.J., Archontoulis, S.V., Confalone, A., Constantin, J., Cuadra, S.V., Debaeke, P., Faye, B., Grant, B., Hoogenboom, G., Jing, Q., van der Laan, M., da Silva, F.A.M., Marin, F.R., Nehbandani, A., Nendel, C., Purcell, L.C., Qian, B.D., Ruane, A.C., Schoving, C., Silva, E., Smith, W., Soltani, A., Srivastava, A., Vieira, N.A., Slone, S., Salmeron, M., 2022. Are soybean models ready for climate change food impact assessments? Eur. J. Agron. 135, 15. https://doi.org/10.1016/j.eja.2022.126482.

Kumar, A., Singh, R., Jena, P.P., Chatterjee, C., Mishra, A., 2015. Identification of the best multi-model combination for simulating river discharge. J. Hydrol. 525, 313–325.

Martre, P., Wallach, D., Asseng, S., Ewert, F., Jones, J.W., Rotter, R.P., Boote, K.J., Ruane, A.C., Thorburn, P.J., CaMAAsrano, D., Hatfield, J.L., Rosenzweig, C., Aggarwal, P.K., Angulo, C., Basso, B., Bertuzzi, P., Biernath, C., Brisson, N., Challinor, A.J., Doltra, J., Gayler, S., Goldberg, R., Grant, R.F., Heng, L., Hooker, J., Hunt, L.A., Ingwersen, J., Izaurralde, R.C., Kersebaum, K.C., Müller, C., Kumar, S.N., Nendel, C., O'Leary, G., Olesen, J.E., Osborne, T.M., Palosuo, T., Priesack, E., Ripoche, D., Semenov, M.A., Shcherbak, I., Steduto, P., Stockle, C.O., Stratonovitch, P., Streck, T., Supit, I., Tao, F., Travasso, M., Waha, K., White, J.W., Wolf, J., 2015. Multimodel ensembles of wheat growth: many models are better than one. Glob. Change Biol. 21, 911–925. https://doi.org/10.1111/gcb.12768.

Motha, R.P., 2011. The impact of extreme weather events on agriculture in the United States. In: Sivakumar, M.V.K., Stefanski, R. (Eds.), Challenges and Opportunities in Agrometeorology. Springer, Dordrecht, pp. 397–407.

Neuman, S.P., 2003. Relationship between juxtaposed, overlapping, and fractal representations of multimodal spatial variability. Water Resour. Res. 39 (8), 1–11. https://doi.org/10.1029/2002WR001755.

Qi, Z., Helmers, M.J., Malone, R.W., Thorp, K.R., 2011. Simulating long-term impacts of winter rye cover crop on hydrologic cycling and nitrogen dynamics for a corn-soybean crop system. Trans. ASABE 54 (5), 1575–1588. https://doi.org/10.13031/2013.39836.

Qi, Z., Ma, L., Bausch, W.C., Trout, T.J., Ahuja, L.R., Flerchinger, G.N., Fang, Q., 2016. Simulating maize production, water and surface energy balance, canopy temperature, and water stress under full and deficit irrigation. Trans. ASABE 59 (2), 623–633. https://doi.org/10.13031/trans.59.11067.

Sándor, R., Ehrhardt, F., Grace, P., Recous, S., Smith, P., Snow, V., Soussana, J.F., Basso, B., Bhatia, A., Brilli, L., Doltra, J., 2023. Residual correlation and ensemble modelling to improve crop and grassland models. Environ. Model. Softw. 161, 105625. https://doi.org/10.1016/j.envsoft.2023.105625.

Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 461–464. https://doi.org/10.1214/aos/1176344136.

Shuttleworth, W.J., Wallace, J.S., 1985. Evaporation from sparse crops – an energy combination theory. Q. J. R. Meteorol. Soc. 111, 839–855. https://doi.org/10.1002/qj.49711146910.

Singh, A.K., 2013. Water and Nitrogen use Efficiency of Corn (*Zea mays* L.) under Water Table Management. McGill University, Montreal, Canada (Ph.D. thesis).

Uzoma, K.C., Smith, W., Grant, B., Desjardins, R.L., Gao, X., Hanis, K., Tenuta, M., Goglio, P., Li, C., 2015. Assessing the effects of agricultural management on nitrous oxide emissions using flux measurements and the DNDC model. Agric. Ecosyst. Environ. 206, 71–83. https://doi.org/10.1016/j.agee.2015.03.014.

Wallach, D., Mearns, L.O., Ruane, A.C., Rötter, R.P., Asseng, S., 2016. Lessons from climate modeling on the design and use of ensembles for crop modeling. Clim. Change 139, 551–564. https://doi.org/10.1007/s10584-016-1803-1.

Wan, Y., Chen, J., Xu, C.-Y., Xie, P., Qi, W., Li, D., Zhang, S., 2021. Performance dependence of multi-model combination methods on hydrological model calibration strategy and ensemble size. J. Hydrol. 603, 127065. https://doi.org/10.1016/j.jhydrol.2021.127065.

Zaherpour, J., Mount, N., Gosling, S.N., Dankers, R., Eisner, S., Gerten, D., Liu, X., Masaki, Y., Schmied, H.M., Tang, Q., Wada, Y., 2019. Exploring the value of machine learning for weighted multi-model combination of an ensemble of global hydrological models. Environ. Model. Softw. 114, 12–128. https://doi.org/10.1016/j.envsoft.2019.01.003.