Response to Peer Review Comments on the Draft Final Report
Estimation of Crop Evapotranspiration in the Sacramento-San Joaquin Delta

Technical Memorandum

Prepared by: Josué Medellín-Azuara, Jesse Jankowski, Kyaw Tha Paw U,
Eric Kent, Jennae' Clay, and Jay R. Lund

University of California, Davis

February 12, 2018

Summary

This memorandum summarizes response to comments provided by the peer review panel on a draft Final Report for the Delta Consumptive Use project. The peer review panel members are: Dr. Richard Allen (University of Idaho, Kimberly), Dr. Richard Snyder, (retired, formerly UC Davis Land, Air and Water Resources), Dr. Thomas Trout (retired, formerly USDA-ARS), and Byron Clark (P.E. and Consultant, Davids Engineering, Inc). The reviewers all provided useful and insightful comments. Their comments were also supportive of the overall effort.

The peer review panel was provided with final drafts of the main report and technical appendices. Dr. Allen and Byron Clark provided inline comments to the draft final report and selected appendices. Dr. Thomas Trout provided both inline comments on the main report and a peer review report with additional information. Dr. Snyder also provided a peer review report of the final report and appendices. Comments from non-peer review panel members or organizations such as the Delta Stewardship Council are not included in this Technical Memorandum, but they were addressed to the extent possible in the Final Report and Appendices.

Comments were generally classified in three tiers. Tier 1 comments required minor text edits or clarifications; only those which required significant additions to the text with respect to the peer-reviewed version are included in this memorandum. Tier 2 comments required additional data analysis or modifications to the presentation of results in the report; these are not discussed in detail herein unless these modifications depart from other suggestions or were not made at all. Tier 3 comments represent outstanding matters which are acknowledged but would require additional time and resources beyond those available for the project to pursue. Thus they are reproduced or summarized in this memorandum with additional discussion from the study team. Comments addressed in this memorandum were generally categorized by topic: Model Methods, Estimates and Presentation of Results, Field Campaign Methods and Results, and Detailed Comparisons between Model and Field Campaign Methods.

Given the broad range of information covered in this study and the diverse participation by many experts, unresolved issues should be expected. Some peer review comments and responses addressed in this memorandum reflect unresolved disagreement regarding the sizable differences between model estimates and field campaign measurements and estimates of ET in the Delta. The field campaign measurements and estimates using deployed equipment showed overall consistency with measurements from independent meteorological station networks such as Fluxnet-Ameriflux and published estimates and measurements from previous studies the Delta. Much of Appendix B is devoted to explaining the data correction, uncertainty analysis, the influence of various reference ET estimates (Spatial CIMIS, Penman-Monteith ETo) on EToF, and the comparison of sensible heat flux (H) and ground heat flux (G) values. Several modeling groups stated that the field estimates seemed lower than would be expected when compared to their estimates and some literature values. Resolving these differences would require time beyond that available for the present study. Further work is needed to better understand and explain differences between ET estimates by remote and field-collected data. The microclimates within the Sacramento-San Joaquin Delta may cause different conditions than those assumed by both remote and field-based estimation methods, though perhaps a simpler explanation will emerge. Future field campaigns may benefit from expanding the network of collaborators to increase transparency and robustness in field measurements and ET estimates.

The discussions in this memorandum also reflect opportunities for research, comparative study, and collaboration on evapotranspiration and the Delta. All data in the report, plus additional intermediate outputs and previous draft data submissions, are available in the project GitHub repository (https://github.com/ssj-delta-cu) for independent analysis. The GitHub repository will be made public following the report's release in early 2018.

Model Methods, Estimates, and Results Presentation

Several comments pertained to the seven models compared in the study, their results (including multiple draft data submissions not included in the Final Report), and the way results were presented. Outstanding comments and responses from the report preparation team follow.

Dr. Trout suggested to clarify the purpose of the report, since the draft title implies is to determine Delta ET and some sections report Delta ET spatial and temporal trends using means from all methods. Furthermore, he considered it appropriate to have two reports; one on general ET results and another one comparing estimation methods.

The title of the report was modified to be more descriptive. The team considers that the purpose is clearly stated in the executive summary and other places: "This report investigates crop consumptive use in the Sacramento-San Joaquin Delta ("Delta") of California using a comparative approach with several prominent methods for estimating crop evapotranspiration, including estimates based on crop coefficients, water balances, energy balance using remote sensing, and field measurements." This report fulfills the purpose of providing estimates from various methods comparatively and investigates the sources of discrepancy between them. Therefore the team considers it unnecessary to produce separate reports to that end.

Dr. Trout indicated methods vary widely in what they measure and how they estimate ET, highlighting the value of the details on each method including: 1) general category (two-step), ground-based, satellite, and hybrid ; 2) spatial scales and extrapolations; 3) temporal scales in measurement; and 4) environments in which ET estimates present larger uncertainties. An expanded Table 2 was suggested.

The team revisited Table 2 on model method descriptions to assure its adequacy given the report sections and the overall objectives of the study. Small edits were made, but it was concluded that further details on models can be found in their respective technical appendices. Substantially expanding Table 2 to accommodate details on estimates, measurements, scales and other model taxonomy may undermine its summarizing purpose given the broader audience of the main report.

Dr. Trout also requested clarification on the contribution of ETa in the non-cropping season, either to the ETa of the summer crop or to fallow. Additional explanation on high EToF values during the winter months was also suggested. Lastly, the separate comparison of methods during the irrigation and non-irrigation seasons was recommended.

Because the land use surveys employed to mask ET results for this study were based on a single crop "snapshot" per year, ETa estimates in the non-cropping season accrued to the summer crop rather than fallow-land ETa estimates. The estimated ET over the water year increases from low in the late fall and early winter to peak in the summer months during the irrigation season for annual crops. High EToF values during the winter in Figure 18 are more evident for the non-remote sensing methods (CalSIMETAW and DETAW). Additional explanations for these outlying values have been added in Section 3.2.4 and elsewhere to explain disparities in EToF during the non-irrigation season. The charts in the report show the range of ET estimates in reasonable detail for both the irrigation season and non-irrigation seasons, and the timing of these seasons could vary considerably by crop and regions of the Delta, so no further breakdown of ET values temporally was pursued. Concluding remarks and discussions of discrepancies have been included in Section 4 of the main report, expanded in Appendix A, and a full set of charts for all crops appears in Appendix K.

Dr. Trout did not fully agree with the assertion that 'a 30x30-meter resolution pixel typically encompasses soil and crops with a variety of root and canopy systems in vertical layers, affecting the accuracy of energy balance calculations for ET estimation that rely on single or *two-layer turbulent transfer approximations and horizontal homogeneity:' "This scale includes sunlit vegetation, shaded vegetation, sunlit soil, and shaded soil, but I wouldn't expect large differences in crop canopies or soil conditions."*

Dr. Paw U suggests that variations in vegetation and soil sun exposure are built-in to the assumptions of remote sensing models. He disagrees with Dr. Trout's view that soil and plant characteristics don't change over short horizontal distances, however. There are edges of soil types that cross fields, according to soil maps of several islands the team reviewed, that would indicate substantial soil changes over short distances.

Dr. Trout requested that the report "clarify how sensitive the various methods are to land classification. I believe most RS methods are relatively insensitive, and several other methods did not use the LIQ data."

Additional clarifications have been added to the report to explain which methods rely on land use data entirely (CalSIMETAW and DETAW) and which ones used either Land IQ or other land use datasets for intermediate data processing and validation. The land use data from Land IQ was largely employed by the study team to summarize ET estimates by method, as explained in the Section 2.3 and in more detail in Appendix A. Hence a sensitivity analysis on land use information was not conducted separately.

Dr. Trout suggested different regrouping of methods for method-to-method comparisons, in particular the comparison of DisALEXI to SIMS.

Appendix A shows a wider range of intercomparisons between methods and to the field-based estimates, with additional statistical tests. This is also presented in the context of modeling and field measurement team discussions over the course of the project. Appendix A has been expanded to address some of Dr. Trout's additional comments in this regard, including additional comparative statistics and individual field-to-model comparisons not included in the peer-reviewed version. Thus the need for additional groupings seems less pressing at this point.

Other general concerns expressed by Dr. Trout have been addressed in the main report or its appendices, including improved Appendix H (UCD-METRIC and Appendix I (UCD-PT) explanations to clarify the use of ETr and ETo and additional methodological issues. Further details on methods should be consulted in the referenced peer-reviewed literature.

Mr. Clark requested an analysis of how and why model results changed and improved since the Interim Report, with increased learning and access to field data: "Did input data or assumptions change? Did modelers refine results based on interim estimates from other models and review of field data?"

The Final Report is focused on presenting each model's best estimate, and the Interim Report already presented the initial "blind comparison" results prior to group learning and the field data release, so further analyses of prior data was not don. The publicly available data from the report's entire history presents an opportunity for future studies such as this.

Mr. Clark noted that the t-tests on differences between ET estimates may be erroneous since the ensemble mean includes the model that is being tested against it. T-tests were done between monthly average estimates for alfalfa corn, and pasture over the DSA between individual models, the ensemble mean, and field data.

This is technically true, but the ensemble mean of models is presented frequently in the Final Report and represents a benchmark against which substantial deviations can be observed. The main report now contains a better explanation of the t-tests and their caveats, and Appendix A describes the process in-depth and plots the full set of inputs and test results. Again, making the data from this project publicly available is an opportunity for more complex statistical tests between datasets.

Dr. Allen: "Can you discuss why the water consumption from fallow lands was so significant? Is it due to evaporation from precipitation events or from possible capillary rise from shallow ground water? The ground-based measurements tended to show very low ET from fallowed fields, albeit, their time frame was short and only a few fields were instrumented. Some discussion would be useful."

Some discussion of fallow land results and model differences is contained within Section 4.2.2 of the Final Report, though the limited scope of field data makes it difficult to determine the significance of these differences. Dr. Allen presents some potential reasons for the higher model estimates, and the upcoming study of fallow fields in the 2018 growing season will provide additional insight on this matter.

Dr. Allen was concerned about the influences of the drought on water supply and related ET from agricultural lands in 2015 and 2016: "It is not clear to me why the impacts of the large drought and on surface water supply are not mentioned. Were drought impacts small this low in the overall Sac/SJ basins?"

Most users in the Delta have very senior water rights, and were therefore much less affected by the drought. The voluntary fallowing agreements executed in 2015 were an exception to this.

Dr. Allen was concerned about potentially low ETo estimates reported by CIMIS: "A brief calculation comparison that I made between ETo computed by the FAO/ASCE PM ETo equation *and as provided by CIMIS for the Holt 248 CIMIS station shows that the CIMIS ETo tends to be about 10% stronger than that of the PM. This would impact the EToF values reported in this study, as they would need to run about 10% lower than literature values that are generally intended to be based on the national/international standardizations (PM). Therefore, overall there is reason to judge that the EToF values, of this report are understated by about 18 to 20%, due to measurement bias and ETo basis, as compared to standard literature values."*

The Spatial CIMIS ETo datasets used by most methods and referenced throughout the report are calculated using the ASCE version of the Penman-Monteith equation (EWRI-ASCE, 2005), which is reported as "PM ETo" rather than "CIMIS ETo" at the individual stations. Therefore EToF computations and other comparisons using ETo values should not be affected by Dr. Allen's concerns. The report text has been expanded to better distinguish and cite these source datasets. Plots in Section 3.2.3 comparing Spatial CIMIS to CIMIS stations do use "CIMIS ETo" values at the stations, but these data were not used in any other part of the report. Dr. Allen's comment may be relevant for irrigators deciding which ETo values from the CIMIS website to use.

Dr. Snyder expressed concerns about the METRIC methods employed the study, such as the estimation of sensible heat flux based on the the slope of the change in temperature, and recommended revisiting their methods. In his opinion, this is more problematic than an operator-based selection of hot and cold pixels.

This potential methodological improvement would require direct interface with various METRIC user groups to discuss alternatives to advance the METRIC model in response to this caveat. Future applications of the model to the Delta or other places may serve as a venue for these discussions.

Field Campaign

Other comments pertained to the specific methods employed by UC Davis for field-based ET measurements, as well as their results reported for the study. Outstanding comments and responses from the field team follow.

Dr. Allen was concerned about the irrigation status of fields in the Delta, particularly related to the pasture stations where field stations were located: "Not so much that the measurement sites might have been under-irrigated (although this needs to be emphasized), but that some population of Delta fields may be under-irrigated, so that derived EToF values should not be expected to be as high as potential values reported in the literature. Can there be an estimation of the percentage of total agricultural fields that may suffer from some water stress? This would be useful, even if labeled as speculative."

As expressed earlier, water stress conditions are rather rare in the Delta, even during droughts, due to seniority in water rights for most Delta farmers. Some descriptive information on the field campaign conditions is presented in Appendix B, and the 2016 field campaign repository (https://github.com/ssj-delta-cu/ssj-field-measurements-2016) includes photographs of the fields where stations were deployed.

Dr. Allen was concerned about the low field-based ET estimates reported: "It seems that all [field] ET measurements should be multiplied by 1.08 to account for under-measurement by the NRLite net radiometers."

Dr. Paw U suggests that the evidence in the literature seems mixed on this issue. Dr. Snyder considers the NRLites to actually overestimate Rn values by a factor of by 1.04. Uncertainty is acknowledged in Appendix B, but a definite bias of 1.08 or another figure is not considered structural by the field campaign team.

Dr. Allen: I recommend that additional teams outside UCD be employed to make field measurements.

Appendix B includes comparisons with other independent investigations and published ET estimates from literature. Therefore this was not deemed necessary at this point, though further field experiments are suggested.

Dr. Allen suggested that further field data collection and cooperation with remote sensing models may not increase agreement "if the field techniques and instrumentation are not changed. Better instrumentation (four-component radiometers) and infrared gas analyzers are highly recommended for future studies. This is a valuable water resource."

The field team considers that inter-radiometer differences, in the range of 0.1 mm/day (Figure B-6.2 in Appendix B), do not represent a major problem for the intercomparison. Individual radiometer calibration seem more problematic, as localized conditions may affect individual leveling and calibration information. This problem also exists for the more sophisticated four-stream radiometers. Dr Paw U mentions that eddy covariance measurements, which do not rely on net radiation, also showed low ETa values. Hence the need for more expensive equipment is questionable.

Dr. Allen expressed concerns and requested clarification on the sonic transducer corrections for field data. He also considered it strange that corrections did not vary with horizontal and vertical wind velocities. Lastly, Dr. Allen mentions other methods and sensors may reduce calibration requirements.

Appendix B now includes additional figures showing the 11% correction found by Kochendorfer et al. (2012) for the RM Young anemometers and the 5% corrections for the CSAT3 anemometers. Since horizontal wind velocities are not employed in direct flux measurements, the field campaign team considers that varying horizontal corrections were not applicable to their measurements. The surface renewal frequently employed in this study requires calibration for practical reasons. More costly sensors may reduce the need for calibration yet increase the time and resources needed to maintain them, which were beyond those allocated to this project.

Mr. Clark: "Was fetch analysis of the field data conducted to ensure that turbulent fluxes (sensible and latent heat flux) were derived from locations within each field? Contributions from outside of the fields could result in underestimation, depending on the conditions adjacent to the fields."

Appendix B describes field conditions in the vicinity of each station. According to the field campaign team, more than adequate fetch was obtained by adequate sensor placement height and location.

Mr. Clark suggested that field measurements have a limited spatial coverage and are also challenged by the need for closure in the energy balance.

Field measurements' limited spatial coverage is inherent and well-recognized in the main report's conclusions. A full description of the closure factor alpha is provided in Section B.6 of Appendix B. The statistics conducted to obtain the closure factor indicate that it is within the range of published work.

Model and Field Intercomparisons

The remaining comments addressed the intercomparison that was conducted between paired methods and between methods and the field data. They were most frequently concerned with the methods employed to make the comparison, they contributed to discussions on the reasons for discrepancies, or they requested additional comparisons be made. Outstanding comments and responses from the report preparation team follow.

Dr. Trout examined the comparisons between field and ET estimates in the context of similar studies in the existing literature. In his view, these comparisons seemed poor at the time of the peer review and a concerning part of the study. In addition, Dr. Trout wanted to see comparisons between models and field data for every day, including model interpolation between overpass dates.

Since the release of the peer review report and appendices, additional comparative timeseries plots were added to Appendix A to show all available data, both on overpass dates and interpolated values, for all fourteen 2016 field stations. Further analyses of daily data beyond overpass dates was not further explored, however, since it would introduce additional discrepancies due to temporal interpolation methods unique to each model. Furthermore, ITRC-METRIC and UCD-METRIC did not supply daily data between overpass dates. Standardization of overpass dates, cloud masking methods, and temporal interpolation methods between models may increase convergence, but modelers typically preferred their own selected methods.

Dr. Allen advised that averaging ET results over a 3x3 pixel grid may be too small of an area: "I would be careful in using such a small area since some methods employ thermal sharpening which will, at a minimum combine information from at least a 3x3 pixel (30 m pixel) area. Some sharpening methods will take information from areas even further away. The specific 3x3 or x by x area will vary among methods. Therefore, it would be best to select large fields to sample so that, at a minimum, a 5x5 or 7x7 (210 x 210 m) area can be sampled and compared… Note that the worst pixel to sample is the one containing the field equipment, since there is generally human impact (trampling, pathways, etc.) plus the reflection off the equipment (what is the albedo of the instrument frame and solar panel?). Furthermore, the flux systems sample LE and H from some distance upwind of the equipment, so that a larger area should be sampled from satellite to be congruent… If field polygons are available, that might be an even better basis for comparison. However, then bleeding effects of thermal off the boundaries can be a problem. You can refer to the following paper that shows some of the field edge effects for L7 and L8: Kilic, A., Allen, R., Trezza, R., Ratcliffe, I., Kamble, B., Robison, C., & Ozturk, D. (2016). Sensitivity of evapotranspiration retrievals from the METRIC processing algorithm to improved radiometric resolution of Landsat 8 thermal data and to calibration bias in Landsat 7 and 8 surface temperature. Remote Sensing of Environment, 185, 198-209." Mr. Clark offered a similar suggestion to average results over the entire field, with an appropriate buffer.

Since these comments, a more detailed comparison of spatial ET averaging scales was conducted to evaluate these concerns. Estimates averaged over 3x3 pixel "grids" were compared to results averaged over the entire farm fields ("parcels," as surveyed by Land IQ) where stations were located. The test showed some small differences at times, but on the whole the two populations were statistically similar and the study team feels that the 3x3 grid is sufficient for result comparison and more representative of the fetch measured by field stations. The main report now references this comparison and full results are included in Appendix A, including spatial results maps to show how each method's estimates varied over each field station parcel on a single overpass day in the summer of 2016.

Dr. Paw U responds that the geometry of the stations is mostly along a single pipe structure, around 3-4 meters tall and less than a 1 meter wide. The surface renewal and eddy covariance H estimates should be representative of an area wider than a single 30x30m pixel. The net radiometer, however, and definitely the G measurements, will have most of its footprint inside one pixel. Dr. Allen may be bringing up possible disturbances of the crop by setting up a station influencing the ET (e.g. some alfalfa got trampled when setting up). This could have an effect, but it wouldn't be for all stations and as he mentions it is probably small (alfalfa might be the biggest effect). The field team generally did not observe much disturbance around the stations. Also, given that the measured Rn value is coming from inside the pixel, including this pixel should tend to make the model results closer to the field estimates anyway.

Dr. Allen requested that model vs. model and model vs. field 1:1 scatter plots give linear equations with a forced (0,0) point.

Slopes and y-intercepts for all possible model-model and model-field comparisons are now provided in figures in Appendix A. The forced (0,0) point was not done, however, since it is very reasonable that one model could give a zero ET estimate while another may not. If this was done, model vs. model biases would increase slightly since the paired y-intercepts averaged 0.4 mm/d. Model vs. field biases would increase more substantially since their y-intercepts averaged about 5 mm/d (models higher than field).

Dr. Allen notes that METRIC models might differ due to software versions: "The ITRC version may be 10 years old. The UCD version may have been a more modern version that was updated for UCD by the University of Idaho within the past four years. METRIC is continually evolving and improving in regard to individual energy balance components and in specific functions for specific crop types, such as trees and vines, when that information is available."

The report text was modified to include this suggestion and indicate the specific METRIC software versions used by ITRC and UCD. This note provides a valuable reminder that ET estimation methods are consistently improving but it is not always easy to migrate changes between models as they are customized to fit the unique experience and skills of their modelers and their applications.

Regarding the assertion that times "where DisALEXI estimated higher ET than SIMS may have been caused by early season irrigation or precipitation events which resulted in evaporation from topsoil," Dr. Allen suggests "it might be useful to show these periods separately, or with different symbols so that the amount of impact is more clearly seen."

These particular differences can be observed in the timeseries plots for each field station which are now located in Appendix A. There are infinitely many ways to present the results of the study depending on the crop, model, time of the year, and location, so the report strived to present the results in the most general form possible. The publicly available results of this study present an opportunity for independent analysis to produce these types of unique comparisons.

Dr. Trout did not feel that presenting R^2 values in model-to-model and model-to-field comparisons was valuable. He suggested that the Mean Bias Error (MBE) may be more useful.

A limited set of comparative statistics were chosen in order to quantify clear trends without duplicating information, and the report team fees that linear regression equations, R^2 values, mean bias, and root-mean-square (RMSE) sufficiently cover these needs. The main report's comparison introduction has been expanded to better explain the significance of each of these statistics, and all resulting values are included in tables in Appendix A.

Dr. Trout requested a direct comparison of CalSIMETAW and DETAW's crop coefficient (Kc) values for each crop to quantify major differences between the models.

This study sought to compare attributes and estimates that were common between as many models as possible, so only ET estimates between these two models were compared. However, Section 4.1.1 does include some insights gained from a detailed examination of assumed crop planting calendars in each model by DWR. Since CalSIMETAW and DETAW both generate tabular values in Microsoft Excel files, the publicly available data from each model that is stored within its respective GitHub repository could easily be independently evaluated at this fine level.

Regarding the suggestion that 'the hourly-to-daily upscaling computation of reference ET that is built into each [METRIC] model (i.e. the use of instantaneous Landsat overpasses data and daily Spatial CIMIS ETo values) may potentially cause systematic differences,' Mr Clark noted: "This is quite interesting, and I look forward to studying Appendix H to better understand, particularly given that future applications of METRIC in California are likely, and ETo is generally preferred for estimation of reference ET."

Since the peer review draft, Appendix H has been expanded to include additional analysis of hourly and daily EToF values to examine some fundamental assumptions of the METRIC model. These comparisons were made exclusively with UC Davis field data collected at eddy covariance stations in alfalfa, corn, and pasture.

Regarding the assertion that, 'due to the unique characteristics in the Delta, with its many channels and wind corridors, regional differences may be significant even over modest distances,' Mr. Clark stated that "it is unclear how these characteristics would affect the constant applied to ETr at the cold pixel for METRIC."

Some concerns were raised by study participants with regard to the potentially narrow range of ET values that could be obtained from the selection of the hot and cold pixels; however, some modeling teams such as ITRC do not consider this to be an issue.

Based on the assumption that the field data was correct, Dr. Snyder recommended comparing instantaneous ET estimates from models, when available, to the corresponding short-term and daily field ET values. Daily method-estimated ET values could also correspondingly be compared to daily field ET values for the length of the study.

It should be noted that the model-to-model and model-to-field comparisons in Section 4 of the Final Report already compare daily field-based estimates to daily ET estimates by models on satellite overpass days. Since the peer review copy of the report was released, additional timeseries plots have also been added to Appendix A which compare daily and overpass ET estimates, where available, from all models to available field data at each station. These plots allow for the visualization of discrepancies in time by method for a specific crop and location. Model results were broadly consistent with field data, but some temporal and site-specific discrepancies merit further examination. This is particularly true for some models in the summer of 2016, but many discrepancies are already addressed in Section 4 of the main report. Most methods did not provide instantaneous ET estimates and none were presented in this report since this high-resolution data was not specifically requested from groups. Available instantaneous estimates submitted by models are available in their respective GitHub repositories.