

Introduction

- Data from the ARM Program include many of the measurements needed by carbon modelers to simulate carbon dynamics in terrestrial ecosystems
- Most models cannot tolerate missing data or gaps
- Missing measurements must be estimated or imputed before the data stream is suitable for use as model input
- Few models can deal with the fine-scale temporal frequency of ARM measurements - data must be statistically aggregated up to larger intervals
- We have produced gap-filled hourly statistical aggregates for selected measurements from 21 ARM SGP facilities for 1996 - 2000 for use in carbon models

Aggregation Methods

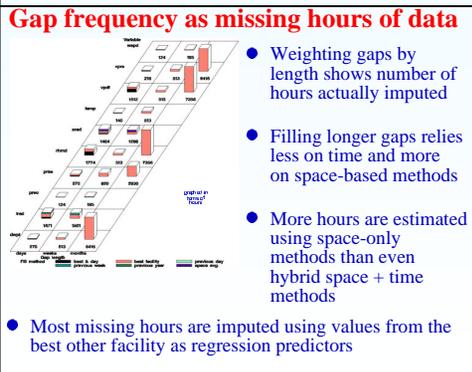
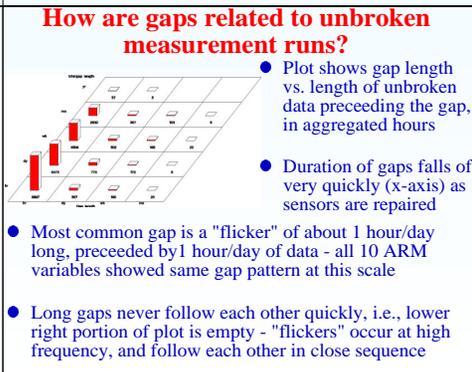
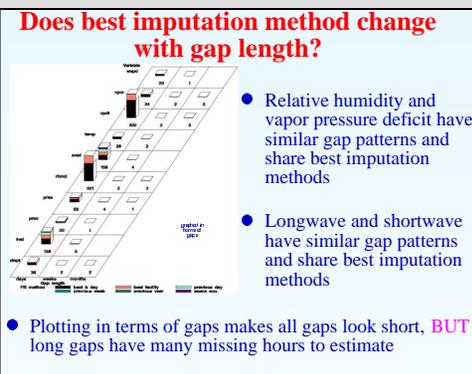
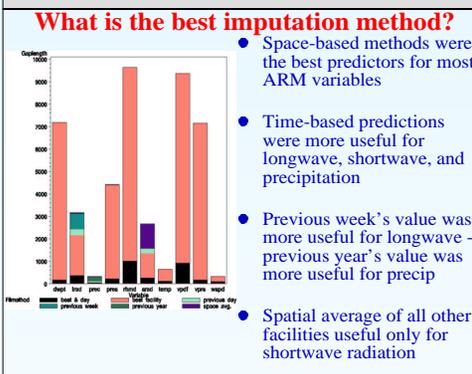
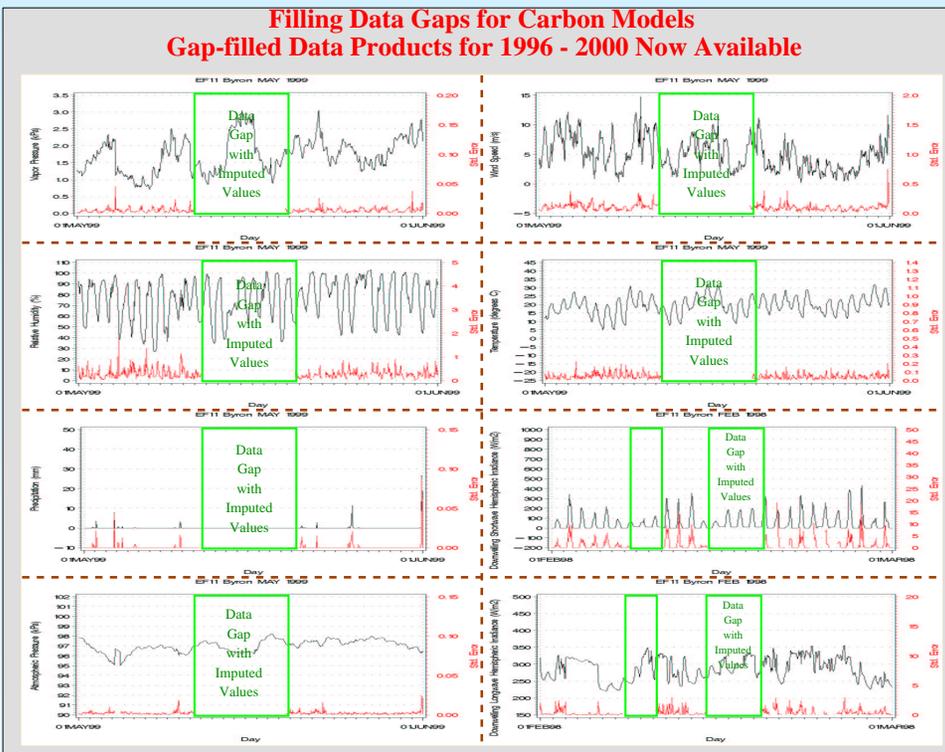
- Ten selected ARM SMOS and SIRS measurements were aggregated up to hourly intervals using means or sums
- Measurements from closest Oklahoma Mesonet site were merged with any SGP facility lacking SMOS
- Measurements were deleted if outside quality control limits describing the range of realistic measurement values, or if Data Quality Reports of problems were filed
- Calculated daylight mask was applied to pyranometers to remove spurious nighttime blackbody radiation values
- Gaps were filled using imputed values to produce clean, complete data sets ready to use in carbon models

Filling Data Gaps

- We developed a generic univariate imputation tool that automatically selects among linear regression models based on values from the same facility at different times and/or different facilities at the same time
- Time-lagged measurements from the previous day, previous week, or previous year, the best alternative facility, and the average of measurements from all other available facilities were used as regression predictors
- Tool selects best regression model, and patches the gaps
- Choice of regression model for each missing hour depends on RMSE and availability of required values - in a 48-hr wide-area-outage, values for best facility or previous day may not be available

Which ARM variables are hardest to impute?

- | Variable | RMSE | R ² |
|--|------|----------------|------|----------------|------|----------------|------|----------------|
| air humidity | 0.07 | 0.92 | 0.07 | 0.92 | 0.07 | 0.92 | 0.07 | 0.92 |
| air temperature | 0.10 | 0.88 | 0.10 | 0.88 | 0.10 | 0.88 | 0.10 | 0.88 |
| precipitation | 0.15 | 0.75 | 0.15 | 0.75 | 0.15 | 0.75 | 0.15 | 0.75 |
| net longwave radiation | 0.20 | 0.65 | 0.20 | 0.65 | 0.20 | 0.65 | 0.20 | 0.65 |
| net shortwave radiation | 0.25 | 0.55 | 0.25 | 0.55 | 0.25 | 0.55 | 0.25 | 0.55 |
| net radiation | 0.30 | 0.45 | 0.30 | 0.45 | 0.30 | 0.45 | 0.30 | 0.45 |
| downwelling longwave radiation | 0.35 | 0.35 | 0.35 | 0.35 | 0.35 | 0.35 | 0.35 | 0.35 |
| downwelling shortwave radiation | 0.40 | 0.25 | 0.40 | 0.25 | 0.40 | 0.25 | 0.40 | 0.25 |
| downwelling longwave + shortwave radiation | 0.45 | 0.15 | 0.45 | 0.15 | 0.45 | 0.15 | 0.45 | 0.15 |
| net longwave + shortwave radiation | 0.50 | 0.05 | 0.50 | 0.05 | 0.50 | 0.05 | 0.50 | 0.05 |
| net radiation + longwave + shortwave radiation | 0.55 | 0.00 | 0.55 | 0.00 | 0.55 | 0.00 | 0.55 | 0.00 |
- Table shows R² and (RMSE) for each variable
 - Parameters differed in predictability
 - Temperature and vapor pressure deficit were easiest to impute - windspeed was intermediate - shortwave harder than longwave radiation - SW showed more use of the long-timelag model
 - Precipitation was hardest - conditional imputation was used - regressions developed only for non-zero precip - if predictor is zero, precip estimate was set to zero - otherwise, precip imputed with best non-zero regression

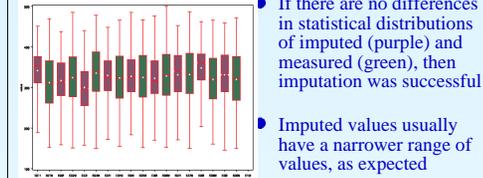


Location of best other facility for imputing



- Precipitation best spatial predictors are at larger scales - offsets nearly the size of ARM CART - a few facilities are the best precipitation predictors for many others
- Precip arrows run perpendicular to rainfall gradient

Comparing distributions of imputed and measured values



- Potential to "tune" regressions to be optimum within the most commonly imputed range - might improve the quality of the imputed estimates
- There are no obvious artifacts from imputation shown by the descriptive statistics for these 10 ARM variables

Conclusions

- No single method of imputation is appropriate for all variables or all gaps - depends on gap length
- No variables have detectable distributional bias inflicted by imputation - no damage by using imputed values
- Precipitation has a "floor" at zero - is better imputed in two separate steps: (1) is it raining? and (2) how much?
- Worst-case may be a variable preferentially imputed by temporal methods for short gaps, in a long gap where they cannot be used, i.e., gaps in longwave >1 week
- Difficulty of estimation could be used to prioritize repair order during multiple outages

Next Steps

- Do a "bad gap" analysis for long, multiple site, multiple variable gaps, i.e., an ice storm scenario
- Perform a "hole punching" experiment to make artificial gaps - retain the actual measurements for a paired test
- Gap-filled hourly data products for these 10 ARM variables for all 21 SGP facilities from 1996-2000 are now available as climatic drivers for carbon models - monthly quicklook previews also available via web
- Join us this Thursday, Dec 11 at 6:15 pm in the Marriott Hotel, Pacific Room A, for an ARM Carbon Reception to learn about these value-added ARM data products
- Bill Hargrove, (865) 241-2748, hnw@fire.esd.ornl.gov, http://www.archive.arm.gov/Carbon