Workflow for QCing and infilling Niwot hourly climate data

1. Hourly data have been aggregated for:
   1. C1
   2. Saddle
   3. D1
2. Hourly data also exist for:
   1. Arikaree
   2. Subnivean Lab
   3. Green Lakes IV
   4. Niwot Ameriflux
3. The following variables are needed to run SNOWPACK
   1. Air temperature
   2. Relative humidity
   3. Wind speed (scalar)
      1. Wind direction can also be used, but is not necessary
   4. Incoming solar radiation
   5. Incoming longwave radiation
      1. To be estimated from T and RH
      2. Compare with LW obs from recent years
   6. Precipitation
      1. Will take daily data and disaggregate
      2. Needs to be gage corrected
      3. Saddle P is too high due to blowing snow
      4. Note: C1 and D1 daily data will come from Tim Kittel’s dataset
         1. Precip will *not* be part of the met dataset
4. Niwot sites have other data that should be QCd and infilled
   1. E.g., soil moisture and temperature, and max/min values of the above data (section 3)
   2. But the SNOWPACK variables are the first priority
5. General
   1. Import hourly C1, Saddle and D1 data into new R workspace
      1. nwt\_hrly\_qc\_all.R
   2. Aggregated, non-QCd data in: ~/Documents/research/site\_data/niwot/nwt\_hrly\_qc/processed\_hrly\_NOQC
   3. 2 hour and 0.5 hour data are not included in this protocol
      1. Data range is from 1988-01-01 (C1, D1) and 1990-01-01 (SDL) to end of data (2014 C1, D1) or 2013 (SDL), depending on station
   4. Datetime is set as MST
      1. Set *tz = “MST”* in R
6. Goals and requirements (Kittel, 2010; p. 5)
   1. Goal: Create a QCed, infilled hourly climate dataset extending back in time as far as possible to analyze changes snowpack energy and snowmelt dynamics
   2. Variables needed by SNOWPACK model: air temperature, relative humidity, wind speed, incoming shortwave and longwave radiation, precipitation
   3. Spatial scale: Points representing the stations (C1, SDL, D1)
   4. Temporal scale: Hourly
   5. Record length: Late 1980s to present
7. **AIR TEMPERATURE**
   1. Run Meek and Hatfield QC protocol on each dataset
      1. Limits
      2. Rate of change
      3. Stationary values
   2. Gap fill through single-station, univariate temporal analysis, then multi-station regression
      1. MicroMet linear interpolation scheme used to gap fill short time spans (e,g, Liston and Elder, 2006; Henn et al., 2012)
         1. Here we rely on the univariate time series from each station:
            1. 1 h missing = linear interpolation between previous and next observation
            2. 2 h to 24 h missing = average of the 24 h before and after each observation
            3. 25 h to 72 h missing = linearly weighted average of ARIMA forecast and backcast generated by the number of observations equal to the gap length before and after missing period

Seasonal ARIMA to account for 24 h diurnal cycle of temperature

In R: *arima(DATA, order = c(1,1,1), seasonal = list(order=c(0,1,1),period=24))*

Consider calculating metrics for different ARIMA orders and time spans (e.g. longer forecasts and backcasts)

* + 1. Regressions made for each month and each 3 h daily time block
       1. Time blocks (each number = hour of day)
          1. 0 ≤ Block 1 < 3
          2. 3 ≤ Block 2 < 6
          3. 6 ≤ Block 3 < 9
          4. 9 ≤ Block 4 < 12
          5. 12 ≤ Block 5 < 15
          6. 15 ≤ Block 6 < 18
          7. 18 ≤ Block 7 < 21
          8. 21 ≤ Block 8
    2. Regressions fitted for:
       1. z ~ x + y (both other stations)
       2. z ~ x (closest station)
       3. z ~ y (next closest station)
    3. Each station therefore has 288 regressions
       1. 12 months X 8 time blocks X 3 station combos
    4. When no stations reporting, use climatological mean to infill
       1. Done per month and time block (96 means)
          1. 12 months X 8 time blocks
  1. V2 of QCed and gap-filled air temperature dataset produced on 2016-05-26
     1. From 1990-01-01 to end of record
     2. For C1, Saddle, and D1
     3. Filled data in column temp\_FILLED2 for each station
     4. Flags are in the flag columns
        1. Flag 0 = missing in original data
        2. Flag 1 = exceeds max/min threshold
        3. Flag 2 = exceeds rate of change threshold
        4. Flag 3 = stuck instrument
        5. Flag 4 = filled with univariate, single-station, temporal infilling
        6. Flag 5 = filled with z ~ x + y regression
        7. Flag 6 = filled with z ~ x regression
        8. Flag 7 = filled with z ~ y regression
        9. Flag 8 = filled with climatological mean for that month and time block

1. **WIND SPEED**
   1. wind\_spd\_scalar is the variable being used
      1. Vector takes direction into account (i.e., wind blowing at 5 m/s from the north for 30 min and then south for 30 min would give a vector average of 0 m/s because the opposite directions would cancel out when averaging)
         1. Scalar does not take direction into account and represents the average of the actual wind speed (i.e., wind blowing at 5 m/s from the north for 30 min and then south for 30 min would give a scalar average of 5 m/s)
         2. Vector average will never be higher than scalar (it can be as high if direction is constant)
   2. QC V1 performed 2016-02-10
      1. Next step is to create regressions and infill
   3. V1 Infilling performed 2012-02-11 using same protocol as air temperature
      1. Issues
         1. C1 is a poor predictor for D1 and SDL, and vice versa
         2. If only C1 is available, wind speeds are too low and have too small a range at D1 and SDL
         3. Some wind speeds are negative (from intercept of regression)
            1. Force to 0?
      2. Consider using an error term to reproduce variability in regressed data
2. **INCOMING SOLAR RADIATION**
   1. V1 QCd and infilled on 2016-02-15
   2. Issues
      1. Using C1 to predict solar at D1 and SDL appears to create a low bias in the maximum and a high bias in the minimum
         1. Regression slopes are less than 1 and intercepts are > 0
         2. Are D1 and SDL cloudier than C1?
      2. Dip in C1 data in 2000s
         1. Instrument error?
   3. Might be more effective to use TOPORAD or a similar protocol to create synthetic data when missing
   4. Or have all regressions be 1:1 with intercept as 0
3. **RELATIVE HUMIDITY**
   1. Data first has to be preprocessed to remove instrument drift
   2. Relative humidity does not hit 100% during every part of the year
      1. This is confirmed looking at the NWT Ameriflux observations
      2. We can estimate monthly maxes by fitting a spline to the data
         1. Spline will then be used to calculate deviation from hypothetical max
            1. Will help identify periods of sensor drift
            2. Then the values will be scaled based on deviation
            3. However, lower RH values do not scale like higher RH values (i.e., greater compression seen in high RH than low)

Scaling will need to be variable based on RH value

High values scaled more than low

* 1. Preprocessing steps:
     1. QC data for max-min threshold (0 and 120)
     2. Find means for monthly RH max based on C1 period of record with newer hygrometers (2010-2014)
     3. Calculate the monthly deviation of observed max from hypothetical max as absolute and relative
     4. The relative deviation becomes the scaling factor for each month
        1. The scaling factor is then scaled by the percentage of the observed relative humidity of the observed monthly max
        2. Meaning low RH values will be scaled less than high RH values
        3. See R code for specifics
     5. Preprocessing looks good for C1 and SDL
        1. Primary issue seems to be reduced variance (the low RH values still increase too much)
           1. Consider changing in future
     6. Preprocessing looks good for D1 except for 1990s period of wonky data
        1. All obs may have to be scrapped during this period.
        2. OR scale preprocessed points by their deviation from minimum.
           1. Might generate proper patterns and variance.
           2. Could confirm against SDL
        3. D1 preprocessed from beginning of record to July 1, 2000
           1. Minimums were scaled to the monthly mean minimum RH using the same protocol as max RH corrections for sensor drift (2016-02-16)
           2. Results not very promising

SDL and D1 observations do not fit well

* + - * 1. May have to be scrapped

Consult with others

* + - 1. NOTE: These data are being omitted and will be gap filled using C1 and SDL
         1. Relationship is too poor with SDL (R2 = 0.17)
         2. Relationship after period is high (R2 = 0.89)
         3. See plots
  1. QC levels 2 and 3 are run on the preprocessed RH data
  2. Infilling performed 2016-02-17
     1. Issues
        1. Infilled data has different spread than observed
        2. Some observations exceed 100%
     2. Solution
        1. Perform regressions on dewpoint temperature and convert that to relative humidity

1. IMPROVEMENTS
   1. Air temperature, wind speed, solar radiation, and relative humidity have all been quality controlled and infilled
   2. The code could use some improvements
      1. Regressions don’t need to be in separate data frames (lists of lists)
      2. The infilling should not loop through every single entry
         1. Think of ways to do this with ifelse statements based on station availability
         2. At the very least the code should be parallelized and/or data should be subset to just the missing obs that need infilling
   3. More robust QC procedures?
      1. Based on other variables
         1. E.g. RH ~ air temperature + solar radiation
      2. Spike detection using the AnomalyDetection package
      3. Go through station reports and flag times where errors were reported by techs
         1. Also note instrument or other important changes
         2. See Kittel (2010) p. 15+ for correction protocols
            1. See p.19 for changing mean offset and variation when instrument changes
         3. Not all inhomogeneities are documented (p. 20), so a reference series would need to be created (e.g., NOAA or Ameriflux at C1)
   4. More robust infilling procedures
      1. Variables aren’t distributed normally so may need to be transformed
         1. Kittel (2010) p. 17
         2. For example, RH data can use arcsin(sqrt(x)) transformation
         3. R2 is not a good performance predictor so estimated and actual values need to be compared (Kittel, 2010, p.18)
      2. Include p-value in regression output
         1. Kittel (2010) recommends rejecting relationship when R2 < 0.6
         2. Include error term in regression to add back noise
      3. Use of a generator like MTCLIM, TOPORAD, etc.
      4. Force y intercept to 0
      5. Use robust regression methods for non-normal, heteroscedastic, autocorrelated data (Kittel, 2010 p. 16)
      6. Kittel (2010) p. 21 says it’s ok to extend station data if the relationship is strong with other station(s)
         1. Could do this for SDL based on D1 so all records extend to late 1980s
   5. C1 is a poor predictor of D1/SDL and vice versa
      1. Use Ameriflux for 1999 and later
      2. B1 (2600 m) is closer in terms of elevation than D1 and SDL
         1. See if hourly data exist