MEMORANDUM

To: John Langstaff and Stephen Graham, U.S. EPA-OAQPS

From: John Hader, Graham Glen, Caroline Foster, Samuel Kovach, Delaney Reilly, Chris Holder, River Williams, Anna Stamatogiannakis, and George Agyeman-Badu, ICF

Date: June 18, 2019

Re: Updates to the Meteorology Data and Activity Locations within CHAD

1. Introduction

In the November 1, 2016 version of CHAD, approximately 18 percent (32,723 out of 179,912) of diary-days are missing values for daily-maximum temperature (Tmax) and thus cannot be used by APEX. The temperature data currently in CHAD originate from a variety of sources, including from the original studies and from EPA or contractors who encoded the study data into CHAD. As discussed in Section 2, we used a methodical process to replace most of these missing values. As part of this exercise, for diary-days without county-location information, we identified county locations for over 10,000 diary-days based on respondent zip code and for over 6,000 diary-days based on the metropolitan locations of several of the studies. Some of the diary-days that received repaired county locations were not missing temperature data; nonetheless, we made the repairs as part of a "cleaning up" of the diary data. After this process, only 0.3 percent (565) of diary-days have missing values for Tmax and remain unusable by APEX.

In the same version of CHAD, six studies have at least 200 minutes per day (on average) of time spent in locations that are not sufficiently clear (they are ambiguous). Unspecified and missing location codes are ambiguous, as are those taking place at a residence or a place of employment without specifying whether they are in the three broad microenvironments (MEs) of indoors, outdoors, or in-vehicle. If studies have an apparent bias (via ambiguity) in time spent in the three broad MEs, then the APEX-modeled exposures will also be biased. As discussed in Section 3, we used paired activity-location information from the other 15 studies in CHAD to derive frequency distributions of location codes used per each activity code, with different distributions intended for reassigning unspecified/missing locations, ambiguous residential locations, and ambiguous workplace locations. For the six targeted studies, for a diary event with an ambiguous location code, we reassigned the location code based on the activity by sampling from these frequency distributions. After this process, the time spent per day in ambiguous locations dropped substantially for the six studies, though one study still had more than 200 minutes per day spent in ambiguous locations. These location-code reassignments will substantially reduce bias in APEX exposure estimates, particularly given that one of the six studies constitutes more than half of all CHAD diary-days.

These modifications do not impact the official EPA CHAD-Master database, which remains unchanged. Instead, the modifications are specific to the version of the diary data used for APEX modeling.

2. Temperature Data

2.1. Overview and Objectives

The current CHAD questionnaire file includes Tmax and daily-average temperature (Tavg; ºF) as well as daily precipitation (inches) and daily number of hours with precipitation. Only Tmax is typically used by APEX modelers, and it is used to help select a set of diaries that have similar temperature values as those experienced by a simulated profile at his/her location on a given modeling day. Diary-days without values for Tmax cannot be selected for use by any simulated profile.

As shown in Table 2-1, approximately 18 percent of diary-days are currently unusable by APEX on the basis of missing Tmax. Less than 1 percent of those are missing all indicators of respondent location (state, county, and zip code) and are not from studies of a single metropolitan area; it will not be possible to identify reasonable temperature data for those diary-days. Most of the remaining diary-days have only state information (no information on county or zip code).

Table 2-1. Information on Diary-days Missing Daily-maximum Temperature Values

			Count	Percent of All Diary-Days	Percent of Diary-days Missing Tmax
Missing Tmax			32,723	18%	100%
→	From the 1980s		14	0.008%	0.04%
	From the 1990s		1,230	0.7%	4%
	From the 2000s		25,512	14%	78%
	From the 2010s		5,967	3%	18%
	Missing All Location Information (state, county, zip code; is not a single-metropolitan study)		111	0.06%	0.3%
	Is a Study of a Single Metropolitan Area		0	0%	0%
	Has State Location but not County (and is not a single-metropolitan study)		30,895	17%	94%
	→	Has Zip Code	30	0.02%	0.09%

Notes: Studies limited to one metropolitan area were put into CHAD without county or zip-code information.
Tmax = daily-maximum temperature

The objective of this task is to use historical meteorological records to identify reasonable temperature values for diary-days currently missing those values. Identifying these values relies on knowing or estimating the geographic location of each diary-day. Since most of the target diary-days identify the respondent's state but not county or zip code, in most cases we have made assumptions about respondent locations within the state.

A structured methodology of identifying appropriate temperature data allows us to identify reasonable temperature values for nearly all diary-days, not just those currently missing temperature data. While we will generally not update temperature data in CHAD that are not already missing (unless we believe the current values are erroneous), we can compare current and "new" temperatures as part of quality control (QC). With this in mind, as detailed in Section 2.2, we developed a hierarchy to assign a county location to nearly all diary-days. Then, as detailed in Section 2.3, we matched county locations to the five closest meteorological stations from the historical records, thus enabling the assignment of temperature values.

2.2. Assigning County Locations to Diary-days

Matching diary-days with nearby meteorological stations requires knowing (or estimating) where the diary-days took place. County is the primary indicator of diary location, though zip codes are also available for some diaries, and assigning temperature data on a county basis is reasonable given the typical spatial resolution of counties and typical temperature gradients.

About 43 percent (77,811) of all diary-days already had county designations. For these diary-days, we "cleaned up" the county names to be more consistent with the names provided by the U.S. Census Bureau. While the county and state locations of diary-days are not used in APEX, creating consistent location designations (and use of the more reliable state-county FIPS designations) made the temperature-assignment process more reliable.

The remaining 57 percent (102,101) of all diary-days had no county locations. As indicated in Table 2-2, 111 had no location information at all and they were not from studies located in a single metropolitan area. We could not assign counties to these 111 diary-days, and thus we could not replace missing temperature data if needed.

Table 2-2. Information on Diary-days Without County Designations

			How County Locations Were Determined (showing counts of diary-days)		
	Count	Percent of All Diary-Days	Metropolitan Study Location	Zip Code	State's Population Distribution
Missing All Location Information (state, county, zip code; is not a single-metropolitan study)	111	0.06%	0	0	0
Is a Study of a Single Metropolitan Area	6,150	2%	6,150	0	0

Has State Location but not County (and is not a single-metropolitan study)	95,840	55%	0	0	84,141
					(14 from 1980s;
					6,139 from 1990s;
					64,046 from 2000s;
					13,942 from 2010s)

→ Has Zip Code	11,699	7%	0	11,635	64
					(1 from 1980s;
					62 from 1990s;
					1 from 2000s;
					0 from 2010s)

Note: Studies limited to one metropolitan area were put into CHAD without county or zip-code information.

For the other 101,990 diary-days without county designations, a small amount (6,150) were from studies located within a single metropolitan area. Diary-days from these studies were originally put into CHAD without county or zip-code information. We made the assumption that all such respondents lived in the primary county associated with the area, as listed below.

- Hamilton County, Ohio for the Cincinnati Activity Patterns Study (CIN)
- Wayne County, Michigan for the Detroit Exposure and Aerosol Research Study (DEA)
- Denver County, Colorado for the Denver, Colorado Personal Exposure Study (DEN)
- King County, Washington for the Seattle Study (SEA)
- District of Columbia for the Washington, DC Study (WAS)

Additionally, a small amount (11,635) of diary-days without county designations had reliable zip codes that we geocoded to their most likely counties, following the process listed below. Note that we used geospatial files representing the year 2000 because most of the CHAD diary-days (129,569 diary-days, which is 72 percent of all diary-days) were from the 2000s, and county boundaries have remained unchanged through the last few decades for nearly all U.S. counties.

- Use GIS software to convert the year-2000 county polygons (from the U.S. Census cartographic boundary files) to centroid points (one centroid per county).
- Use GIS software to identify the county centroid (year 2000) closest to each zip-code centroid (also year 2000; from the zip-code tabulation areas file from the U.S. Census Gazetteer files). These centroid-proximity matches were restricted to within the same state (e.g., a zip-code centroid located in California could only be matched to a county in California).
- A small number of zip codes (145) could not be identified in the Gazetteer files. We identified the county locations of 85 such zip codes with reasonable confidence using Internet searches, leaving 60 zip codes unmatched to counties.

For the remaining 84,205 diary-days without county designations (which includes 64 diary-days that could not be reliably matched to counties via zip code), we assigned them to counties within the state based on population distributions. We used U.S. Census data to calculate the population distributions within each state. Since such distributions change over time, we did this on a decadal basis, covering the decades represented by the CHAD diary-days (the 1980s through 2010s), as indicated below. The majority of such population-based assignments were for diary-days in the 2000s decade (as indicated in Table 2-2).

- **2000s and 2010s:** We queried decadal census data from the U.S. Census Bureau American FactFinder website (filtering by Population Total, the 2010 or 2000 year, and All Counties within United States). The SF1 100% datasets were employed.
- **1980s and 1990s:** We used intercensal data from the U.S. Census Bureau's State and County Intercensal Datasets websites for 1980 to 1989 and 1990 to 1999. The county populations were partitioned by demographics, which we aggregated to county-total population values.

2.3. Assigning Temperature Data to Diary-days

The National Centers for Environmental Information (NCEI) distributes several databases of land-based meteorology station data. We utilized the Global Historical Climatology Network–Daily (GHCND), as it provided QCed daily temperature data at a relatively high spatial resolution across the U.S. We narrowed the GHCND database based on the criteria listed below.

- Stations must be located 24–50º N and 126–66º W (for contiguous U.S.), 51–72º N and 179.999–129º W (for Alaska; we did not use any stations in the far-western Aleutian Islands), and 18.5–22.5º N and 160.5–154.5º W (for Hawaii). Note that these boundaries may extend somewhat into neighboring countries.
- Stations must include Tmax and daily-minimum temperature (Tmin) as typically reported parameters (requiring Tavg was too restrictive; we elected to calculate Tavg as the average of Tmax and Tmin).
- On a decadal basis, stations must report data for the entirety of that decade (or for 2010–2014 for the 2010s).

Some of the GHCND stations were of 'higher quality' than others, as they are part of the U.S. Historical Climatology Network (HCN), the U.S. Climate Reference Network (CRN) and/or the Global Climate Observing System Surface Network (GSN). We preferred data from these stations in our temperature assignments.

In Table 2-3, we indicate the number of meteorological stations per decade, including the number of higher-quality stations, that meet all the selection criteria listed above. In Figure 2-1 and Figure 2-2, for the 1980s and 2010s respectively, we show examples of the geographic spread of meteorology stations (with higher-quality stations differentiated) in North and South Carolina.

Table 2-3. Number of GHCND Meteorological Stations Meeting Selection Criteria, per Decade and U.S. Region

Year	Number of Meteorological Station Counts (higher-quality Stations)[a]		
	Contiguous U.S.	Alaska	Hawaii
1980	6,621 (1,225)	230 (19)	54 (2)
1990	7,207 (1,233)	251 (19)	56 (2)
2000	7,813 (1,151)	341 (21)	72 (2)
2010	8,445 (1,210)	388 (29)	85 (4)

[a] Note that a small number of stations included here may be across the U.S. border in other countries.

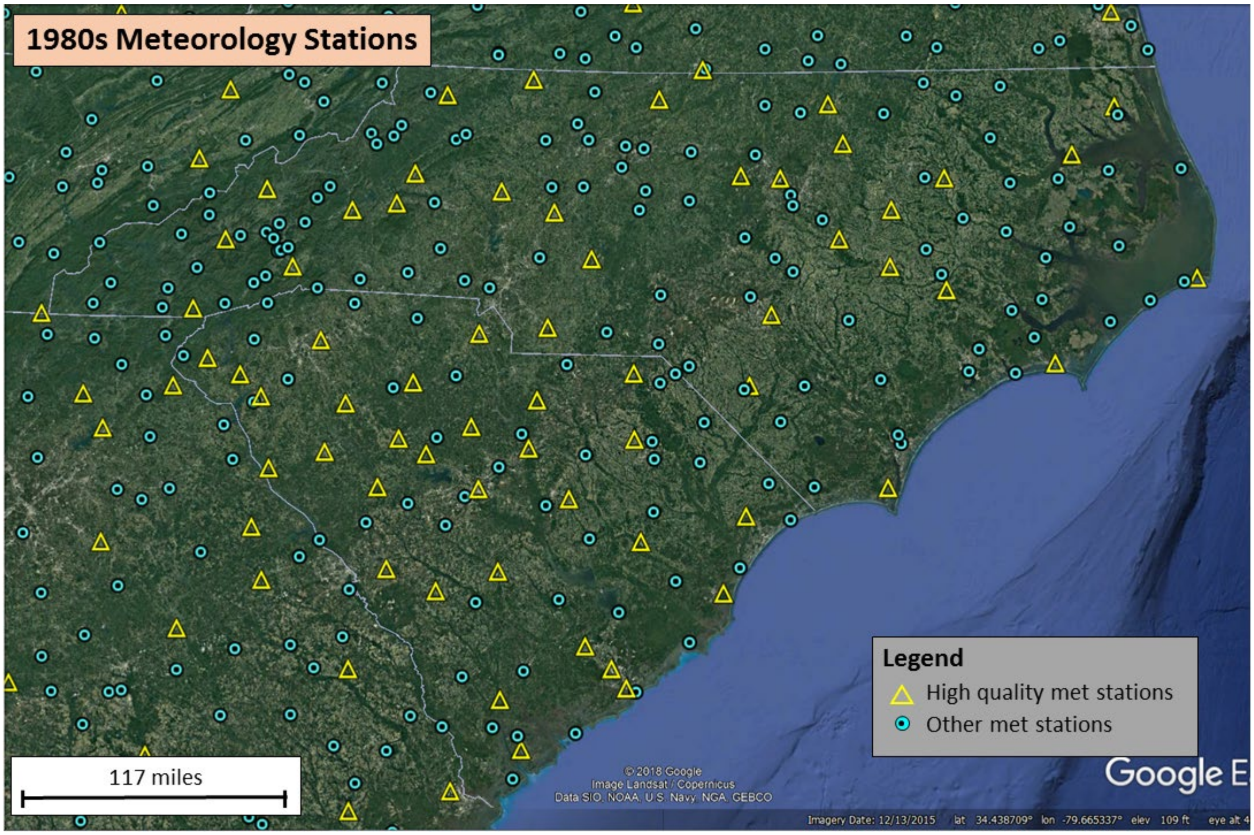
1980s Meteorology Stations
Legend
△ High quality met stations
● Other met stations
117 miles
© 2018 Google
Image Landsat / Copernicus
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Google E
Imagery Date: 12/13/2015 lat 34.438709° lon -79.665337° elev 109 ft eye alt 4

Figure 2-1. GHCND Meteorological Stations from the 1980s Meeting Selection Criteria, in the North and South Carolina Region

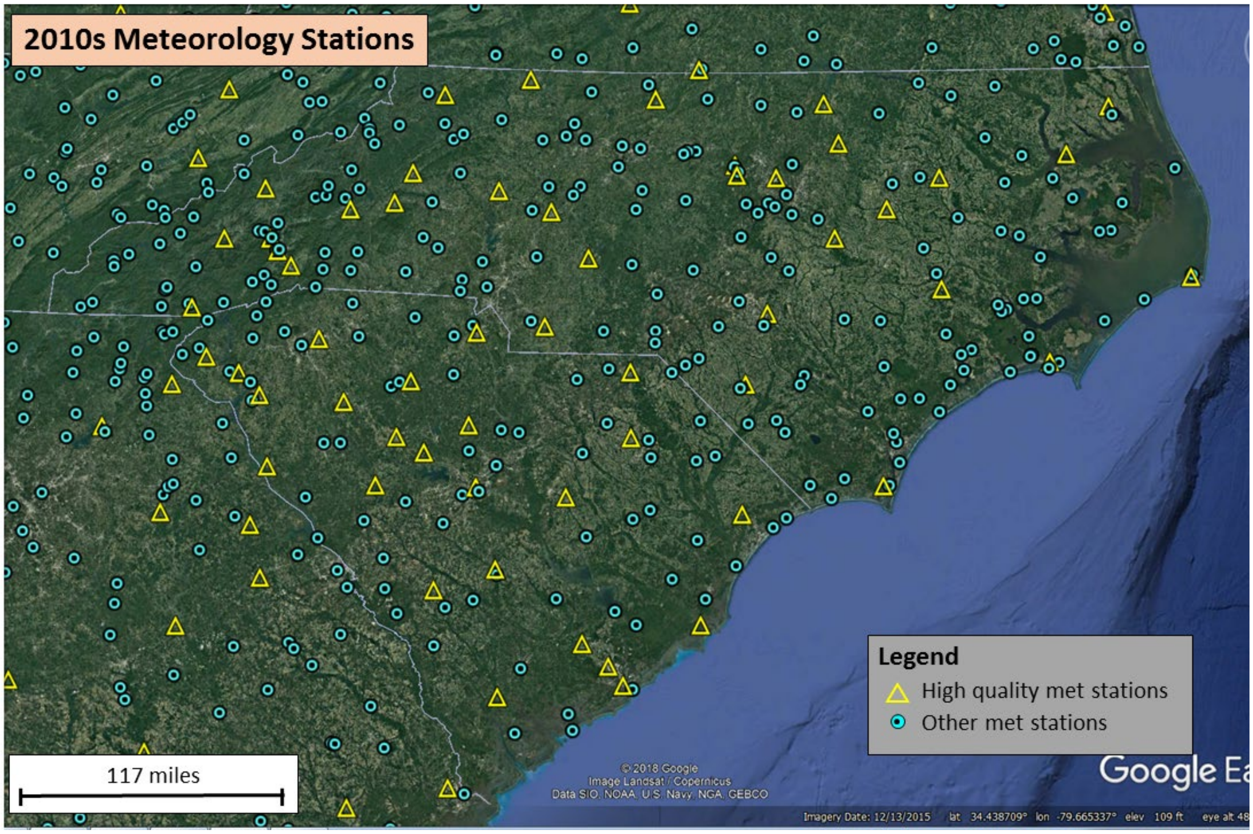
2010s Meteorology Stations
Legend
△ High quality met stations
● Other met stations
117 miles
© 2018 Google
Image Landsat / Copernicus
Data SIO, NOAA, U.S. Navy, NGA, GEBCO
Google Ea
Imagery Date: 12/13/2015 lat 34.438709° lon -79.665337° elev 109 ft eye alt 48

Figure 2-2. GHCND Meteorological Stations from the 2010s Meeting Selection Criteria, in the North and South Carolina Region

By decade (with county locations fixed at the year-2000 definitions), we used ArcMap's "Generate Near Table" tool to map each U.S. county to its five closest meteorological stations from the GHCND dataset. The stations were initially sorted by closest proximity to the county centroid. Then, we resorted the matches to ensure that the closest higher-quality within 30 miles of the county centroid was the preferred station of the five stations.

The median distance from county centroid to the preferred meteorological station was 19 km—only in Alaska were some county centroids more than 100 km from the preferred station, and a few counties in Arizona, California, Nevada, and Texas were 50–70 km from the preferred station. The median distance from county centroid to the fifth selected station was 42 km.

Based on the county location and decade of the diary-day, and the five meteorological stations selected for that county and decade, we identified Tmax and Tmin from the preferred station. If the preferred station's Tmax and Tmin values were missing, then we used the values from the second station, and so on until we identified non-missing values. If none of the five stations supplied non-missing Tmax and Tmin values, then the values were left missing.

Using the method above, 178,893 diary-days (> 99 percent) were matched with new Tmax and Tavg values, leaving 1,019 diary-days (0.6 percent) without matched values. As a QC check, we compared the newly matched temperature values ("new" temperatures) to the existing temperature values where available ("old" temperatures). Using Tmax, there were

146,735 diary-days (82 percent) available for comparison. In Table 2-4, we indicate how many diary-days were negligibly different (≤ 5°), 5–10° different, 10–20° different, or > 20° different.

Table 2-4. Comparison of Old (in Current CHAD-Master) and New (Identified Here) Daily-maximum Temperatures

Difference between Old Tmax and New Tmax	Number of Diary-days	Percent of Diary-days Available for Comparison
≤ 5 °F	101,507	69.2%
5–10 °F	24,604	16.8%
10–20 °F	16,032	10.9%
> 20 °F	4,592	3.1%

During this QC check, we further examined the 4,592 diary-days (3 percent) where the Tmax values were > 20° different. During this step, we discovered that most of these diary-days were from the American Time Use Survey by the Bureau of Labor Statistics (BLS). In 2,431 of the 4,592 diary-days with differences over 20°, they were from the BLS study *and* the old Tmax was equivalent to the old Tavg. This indicated a systematic error in the old BLS temperatures.

Using a similar approach, we compared the old and new Tavg values. The results are indicated in Table 2-5. The results comparing the old and new Tavg values were similar to those for Tmax.

Table 2-5. Comparison of Old and New Average Temperatures

Difference between Old Tmax and New Tmax	Number of Diary-days	Percent of Diary-days Available for Comparison
≤ 5 °F	109,632	74.7%
5–10 °F	24,430	16.6%
10–20 °F	10,271	7.0%
> 20 °F	2,363	1.6%

We further examined the 2,363 diary-days (1.3%) where differences in Tavg values were > 20°. For 1,569 of these diary-days, they were from the BLS study *and* the old Tavg was equivalent to the old Tmax, again indicating a systematic error in the old BLS temperatures.

As an additional check, we examined the mean Tmax and mean Tavg across all diary-days. The mean Tmax and mean Tavg for the old values were 68.0° and 58.4°, respectively. For the new data, the mean Tmax and mean Tavg were 68.4° and 57.8° respectively. The consistency between the two was expected and provides additional assurance.

At the direction of EPA, and given the errors found in the temperatures of the BLS study, we developed a diary dataset using a combination of the old and new temperatures. To create this dataset, we replaced all the old temperatures (maximum and average) of the BLS diary-days. Next, we replaced all previously missing values where new values were available (across all studies). Following these rules, we replaced values for 125,581 diary-days, such that the new diary dataset now has Tmax and Tavg values for 179,347 diary-days. Temperatures remain missing for 565 diary-days, while 53,766 diary-days retained their old temperatures.

In addition to the new temperature data, we updated the dataset with information that was used as intermediate to this process, with fields indicated in Table 2-6.

Table 2-6. Updated or Added Fields in the CHAD Dataset

Field Name	Description
county	Values updated to include newly georeferenced data
state	Values updated to include newly georeferenced data
FIPS	Field added to provide a unique ID to every state-county
old_avgtemp	Field renamed to identify the temperatures (°F) in the November 2016 CHAD
old_maxtemp	Field renamed to identify the temperatures (°F) in the November 2016 CHAD
FIPSfromZip	Field added: TRUE or FALSE—if the county originally was missing, did we identify by zip code?
FIPSfromStudy	Field added: TRUE or FALSE—if the county originally was missing, did we identify by study location?
FIPSfromCountyRandom	Field added: TRUE or FALSE—if the county originally was missing, did we identify by county population distributions in the state?
new_avgtemp	Field added to provide new temperatures (°F) queried in this task
new_maxtemp	Field added to provide new temperatures (°F) queried in this task
ReplacedMaxTemp	Field added to provide the final temperatures (°F) to use in future applications (either the old or new value, depending on the study and other criteria as discussed in this memorandum)
ReplacedAvgTemp	Field added to provide the final temperatures (°F) to use in future applications (either the old or new value, depending on the study and other criteria as discussed in this memorandum)

3. CHAD Activity Locations

3.1. Introduction

Each diary-day reports a series of “events” covering 24 hours. Event durations vary, but each event has one location code and one activity code. To use diaries in APEX, the location codes are mapped to APEX MEs, each of which has a method for determining its air quality. While the number of MEs is flexible, generally all APEX runs distinguish between time spent in three basic MEs: indoor, outdoor, and in-vehicle. Yet six of the location codes are ambiguous, even at that coarse level of defining MEs (i.e., they do not distinguish between the three basic MEs). CHAD is composed of 21 originally separate studies, and some of these studies use these ambiguous codes, but others do not.

These six ambiguous location codes are shown below, and in Table 3-1 we show the average amount of time spent in ambiguous locations (by study).

- Residential:
  - 30000 (Residence, general)
  - 30010 (Your residence)
  - 30020 (Other’s residence)
- Workplace:
  - 33400 (At work: no specific location, moving among locations)

- Unknown:
  - U (Uncertain)
  - X (Missing)

Table 3-1. Average Amount of Ambiguous Time by Study

Study	Average Ambiguous Time (minutes per day)
BAL: Baltimore Retirement Home Study	3
BLS: American Time Use Survey (ATUS), Bureau of Labor Statistics	**498**
CAA: California Adults Activity Pattern Studies	67
CAC: California Children Activity Pattern Studies	0
CAY: California Youth Activity Pattern Studies	101
CIN: Cincinnati Activity Patterns Study	2
DEA: Detroit Exposure and Aerosol Research Study	**1,186**
DEN: Denver, Colorado Personal Exposure Study	16
EPA: EPA Longitudinal Studies	**333**
ISR: Population Study of Income Dynamics I, II, III	58
LAE: Los Angeles Ozone Exposure Study: Elementary School	34
LAH: Los Angeles Ozone Exposure Study: High School	2
NHA: National Human Activity Pattern Study: Air	18
NHW: National Human Activity Pattern Study: Water	18
NSA: National-scale Activity Study	154
OAB: RTI Ozone Averting Behavior Stud	121
RTP: RTP Particulate Matter Panel Study	**1,081**
SEA: Seattle Study	**1,205**
SUP: Study of Use of Products and Exposure-related Behaviors	**804**
VAL: Valdez Air Health Study	2
WAS: Washington, DC Study	16

Note: Bolded studies have relatively large average amounts of ambiguous time.

APEX assigns MEs based only on the location code (not the activity code), and furthermore, APEX uses a deterministic mapping (that is, the same location code maps to the same ME throughout that APEX run). But this rule may lead to an unavoidable bias if applied to certain diary studies. We examined the CHAD activity code that is paired with each location code (on the event level), to determine the likely place of occurrence of each event. Since this is not always a certainty, part of this exercise is to probabilistically assign specific locations to events with ambiguous location codes, based on the paired activity.

3.2. Methods

The starting point is the November 2016 version of CHAD. It has 179,912 diary-days. Two of those (EPA002171 and EPA002172) have been deleted because they each contained 24 hours of missing data.

For our purposes, we divided all location codes into six general MEs and temporarily related them to the location codes shown as shown below, which are unambiguous. The codes are typical examples of the categories shown. For example, 31110 is a car; while not all vehicular travel is in a car, it is reasonable that the air quality in a car would be similar to that found in other types of vehicles.

- IH (indoors at a residence) → Code 30120 (Your residence, indoor)
- IO (indoors elsewhere) → Code 32000 (Other, indoor general)
- OH (outdoors at a residence) → Code 30200 (Residence, outdoor)
- OV (outdoors near traffic) → Code 35200 (Public garage / parking lot)
- O (outdoors elsewhere) → Code 35000 (Other outdoor, general)
- V (in an enclosed vehicle) → Code 31110 (Motorized travel by car)

The six ambiguous location codes had more than one mapping option for a location category, as shown below. They were reassigned location codes based on activity (and occupation where applicable), as discussed later.

- Codes 30000 (residence, general), 30010 (your residence), 30020 (other's residence)
  - Could be either IH or OH; occasionally V or OV
- Code 33400 (at work; no specific location, moving among locations)
  - Could be any, but depends on occupation
    - Occupation TRANS (transportation and material moving)
      - V (specifically 31120, travel by truck)
    - Occupation FARM (farming, forestry, and fishing)
      - O
    - Occupation HSHLD (private household)
      - IH
    - Activity code ≥ 18000 (travel)
      - V
    - Activity codes 17700–17823 (active-leisure activities; exercise activities)
      - OV
    - All others
      - IO
- Codes U (uncertain), X (missing)
  - Could be any

For analysis purposes, we divided CHAD into two parts. The "bad" part consisted of the six studies with at least 200 minutes per day on average spent in ambiguous locations (see Table 3-1; the studies were BLS, DEA, EPA [EPA Longitudinal Studies], RTP [RTP Particulate

Matter Panel Study], SEA, and SUP [Study of Use of Products and Exposure-related Behaviors]). The “good” part consisted of the 15 studies with an average of fewer than 200 minutes per day of ambiguous time.

For the purposes of replacing location codes U and X in the “bad” part of CHAD, we analyzed the “good” part to determine the time fractions in each of the six location categories for each activity code (except activity codes U and X). We excluded any time in ambiguous locations. For example, the “eating” code (14400) divided as IH = 76 percent, IO = 21 percent, OH = 2 percent, O = 1 percent, and OV and V = less than 1 percent. A few activity codes did not have examples in the “good” part of CHAD, and so we mapped them to similar activities. These cases occurred extremely rarely in the “bad” part of CHAD, as well. The number of such cases increased if we stratified CHAD by age group, and for most activities the allocation to the six location categories was not very different between age groups. Therefore, we did not treat age groups separately. We linked the time-fraction distributions to the activities in the six studies in the “bad” part of CHAD. We reassigned U and X locations by activity (excluding activity codes U and X), following these distributions from the “good” part of CHAD.

For the purposes of replacing ambiguous residential location codes (30000 – Residence, general; 30010 – Your residence; and 30020 – Other’s residence), we made separate time-fraction determinations (also from the “good” part of CHAD) where we generally restricted time to three categories: IH, OH, and OV. We used the last of these (OV) for time in the garage or working on cars. We made an exception for selected travel activity codes over 18000, which indicate that the person was in a vehicle. For example, we assigned 18031 (drive a motor vehicle) and similar codes to V. We linked these refined time-fraction determinations to the activities in the six studies in the “bad” part of CHAD, for all events with location codes 30000, 30010, or 30020. We reassigned these locations by activity (for activities other than U and X), following these distributions of time spent. We made an exception for the DEA study, where it was clear that the residential codes up to 30020 were used only for indoor events. Note the before the location reassignments, the DEA study averaged 83 minutes in OH locations but only 29 minutes in IH locations.

In many cases, the same diary had the same activity code for several consecutive events with ambiguous location codes. For example, the person might be sleeping for several hours, but the location is not clear. It would not make sense for them to be relocated part way through, so for such consecutive events we determined the reassignment (from the activity’s distribution across the six location categories) only for the first of such events, and then subsequent events received the same new location reassignment.

3.3. Discussion

As shown in Table 3-2, five of the six studies where we reassigned location codes now have fewer than 200 minutes per day of ambiguous location time. The exception is the SUP study, in which most diaries were shorter than 24 hours and were padded with missing activities and locations to fill out the day. Many of the SUP diaries were previously rejected by APEX, and might continue to be, but most of the other diaries will now be acceptable. In particular, the BLS diaries constitute more than half of CHAD, and they have gone from 498 ambiguous minutes to just 10 such minutes per diary-day.

Table 3-2. Minutes per Day in the Six Location Categories, Before ("Old") and After ("New") Location Reassignments, For the Six Studies With 200 Minutes per Day or More of Time Spent in Ambiguous Locations

	BLS		DEA		EPA		RTP		SEA		SUP	
Location Category	Old	New	Old	New	Old	New	Old	New	Old	New	Old	New
IH	754	1,049	29	1,157	677	903	90	973	0.04	1,121	327	787
IO	79	228	48	95	246	346	131	170	139	145	175	176
OH	22	47	83	83	50	55	36	77	16	73	22	47
O	17	23	19	19	23	23	17	17	24	25	45	45
OV	0.3	1.7	3.3	3.4	24	24	5.8	6.8	1.0	2.1	5.0	5.1
V	70	81	72	72	87	87	80	80	54	54	61	61
Ambiguous	**498**	**10**	**1,186**	**10.3**	**333**	**2.4**	**1,081**	**116**	**1,205**	**21**	**804**	**317**
Indoor Total	833	1,277	78	1,252	923	1,249	220	1,143	139	1,265	503	963
Outdoor Total	39	72	105	106	96	102	59	101	41	99	72	98

Several questions remained, as listed below. We discussed these questions with EPA in May 2019, with decisions noted below.

1. Should the "good" part of CHAD be defined differently?
   a. No, keep it as-is.

2. Should other location codes be deemed ambiguous?
   a. Not at this time.

3. Should this method be applied to the ambiguous events in "good" CHAD?
   a. No.

The last question is perhaps the most important. The CAY, NSA, and OAB studies average over 100 minutes of ambiguous time per diary, which is significant. The same method could be applied there, and might significantly reduce the ambiguous time in those studies. One reason not to apply this method is that the time percentages would then be applied to some of the same studies used to derive the percentages, and this presents the appearance of circular reasoning. It is not exactly circular because we excluded ambiguous time when deriving the percentages, but even so, there may be a correlation between the choice of location code and choice of activity code within a single study. For example, there may be a reason particular to the given study for why some eating events were assigned specific location codes, and others were assigned location X. Hence, it is not clear whether general percentages for all eating events should apply to those (relatively few) coded with location X. This is less of a concern when most or all eating events are paired with location X.

4. Diagram of Processing

In Figure 4-1, we indicate the input and output files for the temperature and location-code updates discussed above, as well as the processing programs and ancillary files. We briefly discuss these files and programs below the figure.

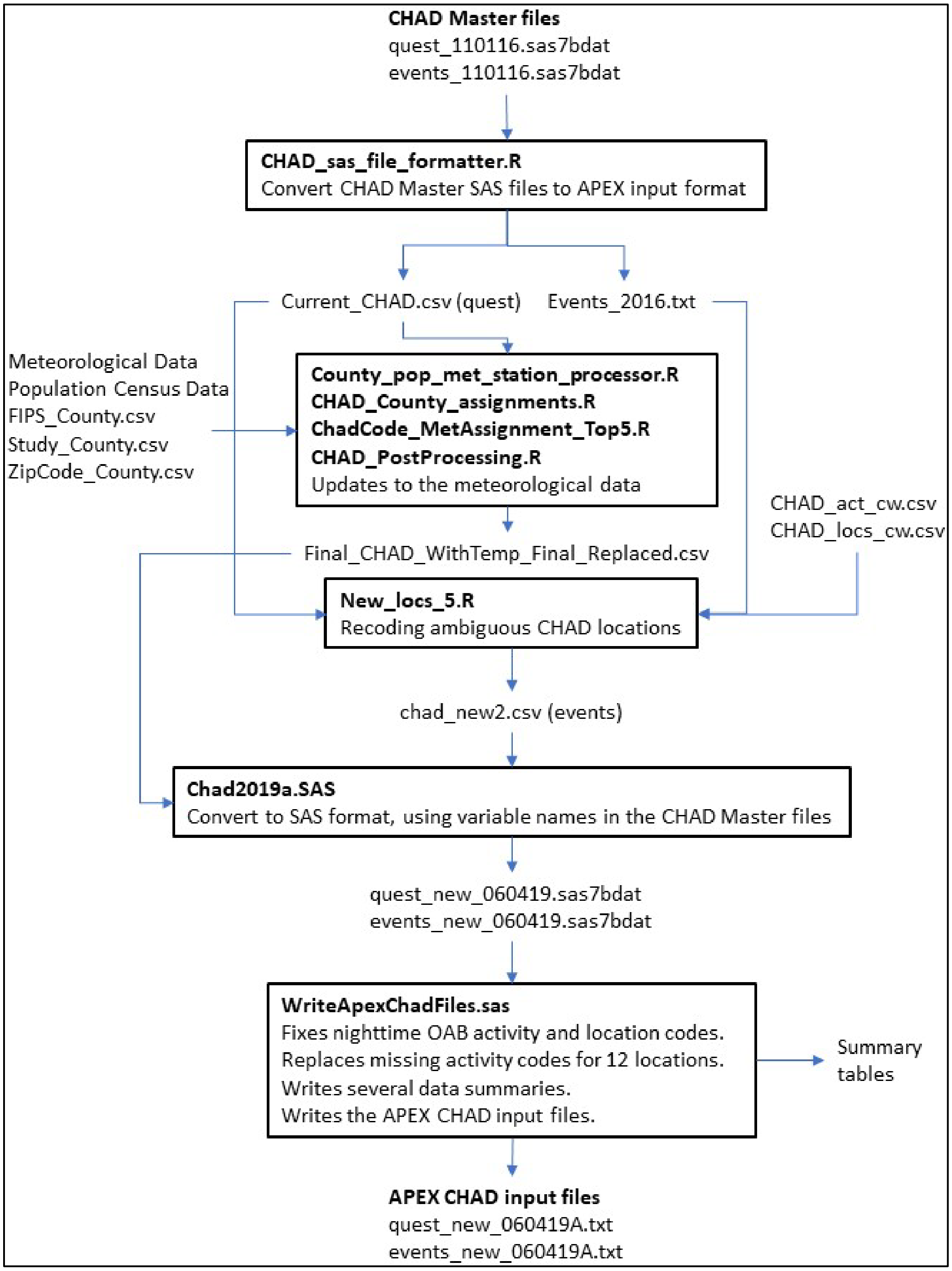
CHAD Master files
quest_110116.sas7bdat
events_110116.sas7bdat
CHAD_sas_file_formatter.R
Convert CHAD Master SAS files to APEX input format
Current_CHAD.csv (quest)
Events_2016.txt
Meteorological Data
Population Census Data
FIPS_County.csv
Study_County.csv
ZipCode_County.csv
County_pop_met_station_processor.R
CHAD_County_assignments.R
ChadCode_MetAssignment_Top5.R
CHAD_PostProcessing.R
Updates to the meteorological data
CHAD_act_cw.csv
CHAD_locs_cw.csv
Final_CHAD_WithTemp_Final_Replaced.csv
New_locs_5.R
Recoding ambiguous CHAD locations
chad_new2.csv (events)
Chad2019a.SAS
Convert to SAS format, using variable names in the CHAD Master files
quest_new_060419.sas7bdat
events_new_060419.sas7bdat
WriteApexChadFiles.sas
Fixes nighttime OAB activity and location codes.
Replaces missing activity codes for 12 locations.
Writes several data summaries.
Writes the APEX CHAD input files.
Summary
tables
APEX CHAD input files
quest_new_060419A.txt
events_new_060419A.txt

Figure 4-1. Files and Processing Programs Used in this Task

Both the temperature and location-code tasks began with the November 2016 version of the CHAD-master files (*quest_110116.sas7bdat* and *events_110116.sas7bdat*), which we converted to text or CSV files (*Current_CHAD.csv* for the questionnaire file; *Events_2016.txt* for the events file) for easier processing in R programs.

We used four different R scripts to modify temperatures and county designations in the questionnaire file. *County_pop_met_station_processor.R* reformatted GIS data, outputting the ranking of up to five meteorology stations for every county, by decade and reorganized based on distance and station quality. *CHAD_County_assignments.R* filled in missing location data, based on zip code, study, and random assignment based on population density. *ChadCode_MetAssignment_Top5.R* combined the outputs of the previous two scripts to assign temperatures (and other intermediate details) the questionnaire file. *CHAD_PostProcessing.R* cleaned the data of unnecessary fields and reformatted the data for processing back into a SAS dataset. The resulting updated questionnaire file was *Final_CHAD_WithTemp_Final_Replaced.csv*.

The location-code reassignments were made by *New_locs_5.R* (where 5 is the version number of the script). The output events file was *chad_new2.csv*.

The new questionnaire and events files were not directly suitable as input to APEX because they contains extra variables, including both the old and new location codes, details about county reassignments and meteorological stations, etc. The program *Chad2019a.sas* converted the files to SAS format and utilized field names conforming to those of CHAD-Master, producing *quest_new_060419.sas7bdat* and *events_new_060419.sas7bdat*.

Finally, the EPA WAM's program (*WriteApexChadFiles.sas*) processed the above-mentioned SAS datasets in various ways, most importantly producing the APEX-ready diary files (*quest_new_060419A.txt* and *events_new_060419A.txt*).