**Identificar, atribuir y superar la calidad de los datos comunes. problemas de observaciones de estaciones tripuladas**

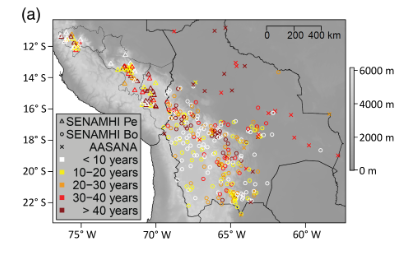
Resumen: Las observaciones climatológicas in situ son esenciales para estudios relacionados con tendencias climáticas y eventos extremos. Sin embargo, en muchas regiones del mundo, los registros de observación se ven afectados por una gran cantidad de problemas de calidad de datos. Evaluar y controlar la calidad de dichos conjuntos de datos es un aspecto importante, a menudo ignorado, de la investigación climática. Además de analizar los datos de medición, los metadatos son importantes para una evaluación integral de la calidad de los datos. Sin embargo, los metadatos a menudo faltan, pero en parte pueden reconstruirse mediante acciones adecuadas, como las inspecciones de la estación. Este estudio identifica y atribuye los problemas de calidad de datos comunes más importantes en los conjuntos de datos de temperatura y precipitación de Bolivia y Perú. Los mismos errores o similares se encuentran en muchas otras redes de estaciones predominantemente tripuladas en todo el mundo. Los observadores pueden rastrear una gran parte de estos problemas hasta los errores de medición. Por lo tanto, la forma más efectiva de prevenir errores es fortalecer la capacitación de observadores y establecer un procedimiento de control de calidad (QC) casi en tiempo real. Los enfoques habituales de control de calidad apenas detectan muchos de los problemas comunes de calidad de datos. Sin embargo, la visualización de datos es una herramienta eficaz para identificar y atribuir esos problemas y, por lo tanto, permite a los usuarios de datos corregir errores y decidir qué propósitos no se ven afectados por problemas específicos. El aumento resultante en registros de estaciones utilizables es particularmente importante en áreas donde las redes de estaciones son escasas. En tales redes, la selección y el tratamiento adecuados de series temporales basadas en un procedimiento integral de control de calidad (QC) pueden contribuir a mejorar la homogeneidad de los datos más que los métodos de homogeneización de datos estadísticos.

Introducción: La evaluación de las tendencias y los extremos climatológicos requiere datos que cumplan con los estándares de alta calidad. Si los datos se ven afectados por problemas de calidad, una consecuencia puede ser resultados engañosos y erróneos. Para eliminar errores, se debe aplicar el control de calidad (QC) antes de analizar registros de observación. La Organización Meteorológica Mundial (OMM) proporciona pautas sobre el control de calidad en varios documentos. Los Servicios Meteorológicos Nacionales y los proyectos de investigación utilizan una amplia gama de enfoques de control de calidad diferentes. En algunos casos, solo se aplican algunos métodos básicos que pueden no garantizar la calidad deseada de los datos. Según Štepánek y otros, falta una metodología general aceptada para el control de calidad. Además, si los datos se recuperan de un conjunto de datos estándar, se consideran adecuados los datos para el propósito previsto a menudo y no se cuestiona.

Aunque los datos satelitales y de reanálisis encuentran un uso generalizado en la ciencia climática, no son un sustituto de las observaciones terrestres debido a limitaciones tales como una menor precisión, longitudes de registro más cortas, problemas de calibración y la necesidad de datos in situ para la validación. Por lo tanto, deben abordarse los problemas de calidad en los registros de observación. Los problemas de calidad de los datos pueden ocurrir en cualquier punto de la cadena de generación de datos, por ejemplo, debido a la configuración o ubicación inadecuada de la estación, mantenimiento deficiente de la estación, lectura errónea del instrumento o digitalización de datos y postprocesamiento imprecisos. Las características y frecuencias de los errores pueden variar entre las diferentes redes dependiendo de factores como el grado de automatización, los instrumentos utilizados, la capacitación de los observadores, el mantenimiento de la estación y el tipo de clima. Por lo tanto, un enfoque de control de calidad puede ser adecuado para una red pero no para otra. No todos los problemas de calidad de datos son detectables controlando los datos de medición. Los metadatos son otra fuente importante en la evaluación de la calidad. De acuerdo con los principios de monitoreo del Sistema Mundial de Observación del Clima, los metadatos (es decir, los detalles y el historial de condiciones locales, instrumentos, procedimientos de operación, algoritmos de procesamiento de datos, etc.) deben documentarse y tratarse con el mismo cuidado que las mediciones. Los metadatos también son necesarios para garantizar que los usuarios finales puedan extraer conclusiones precisas de sus análisis, y desempeñan un papel importante en la detección y corrección de las incongruencias y son esenciales para administrar redes y sistemas de soporte. Sin embargo, en muchas redes en todo el mundo, los metadatos son fraccionados, de baja calidad, o simplemente no están disponibles, por lo tanto, se deben tomar medidas para abordar esto defecto. Asegurar la calidad de un conjunto de datos es más difícil para redes dispersas que para redes densas. Las pruebas de control de calidad importantes dependen de la disponibilidad de estaciones vecinas adecuadas. Además, los registros en redes densas pueden excluirse del conjunto de datos sin perder mucha información, lo que permite eliminar series de tiempo que están incompletas o cuya calidad es incierta. Sin embargo, esto puede no aplicarse a conjuntos de datos de red dispersos. Para obtener el mayor beneficio de las observaciones climáticas en regiones de baja densidad de estaciones, la identificación de errores específicos y, si es posible, las correcciones de errores son particularmente importantes.

En este estudio, investigamos problemas de calidad de datos en observaciones diarias de temperatura máxima/mínima y precipitación en Bolivia y Perú como casos ejemplares para redes de estaciones predominantemente tripuladas. Según Valille y otros, los problemas de calidad de los datos y la falta de observaciones meteorológicas a largo plazo obstaculizan el análisis de tendencias confiables en el área andina central. Lo mismo se aplica a otras regiones, como América Central o África oriental. Analizar las observaciones climáticas del área andina central es un desafío, ya que la región se caracteriza por un terreno complejo y fuertes gradientes climáticos. En consecuencia, los impactos del cambio climático varían. Esto requeriría investigaciones en escalas espaciales relativamente pequeñas, pero las redes de estaciones son escasas en la región. Además, los registros de metadatos son fragmentarios o faltan por completo. Por lo tanto, los registros de estaciones bolivianas y peruanas son buenos estudios de caso para abordar diversos desafíos en el monitoreo del clima. Este estudio describe la situación actual de las redes de estaciones en Bolivia y el sureste de Perú (Sección 2). Se describen las condiciones de medición, las prácticas y la instrumentación, y se describen los esfuerzos para el control de calidad y la recopilación de metadatos. Como los metadatos faltan en gran medida, sugerimos estrategias para reconstruir y generar metadatos (Sección 3). La estrecha colaboración de varias organizaciones, extensas investigaciones en el sitio y análisis detallados de datos nos permitieron caracterizar errores comunes encontrados en los datos e identificar fuentes de error (Sección 4). Con el fin de describir el impacto en los resultados, mencionamos, para cada problema importante de calidad de datos, los índices de cambio climático particularmente afectados y no afectados. Finalmente, este estudio muestra métodos para la detección de errores y discute cómo los registros de la estación pueden usarse potencialmente para ciertos propósitos a pesar de los problemas de calidad existentes (Sección 5).

2. Datos: Con el fin de investigar problemas comunes de calidad de datos de observaciones de estaciones tripuladas, se analizan datos del área central andina. El área de estudio incluye Bolivia y los distritos peruanos Puno, Cuzco y Junín. (Figure 1(a)).

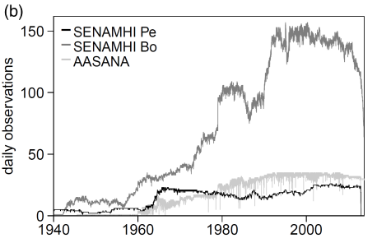


A continuación, los "datos peruanos" se refieren a datos de estos distritos. Sin embargo, tenga en cuenta que los datos de los distritos peruanos utilizados en este estudio no están completos y no cubren el área completa de los distritos. Las variables climáticas investigadas son la temperatura máxima (TX), la temperatura mínima (TN) y la precipitación (PRCP), que son variables clave en climatología (Plummeret al., 2003) y son los parámetros medidos con mayor frecuencia en el área central andina. Se accedió a los datos entre 2014 y 2016 del Servicio Nacional de Meteorología e Hidrología (SENAMHI) Bolivia, el Proyecto Programa Piloto para la Resiliencia al Clima (PPCR) para Bolivia y SENAMHI Peru. El conjunto de datos de PPCR incluye transcripciones que se originan en SENAMHI Bolivia y la administración del aeropuerto civil (AASANA). Existen varios conjuntos de datos digitalizados diferentes que se basan en observaciones climáticas de SENAMHI Bolivia. Son el resultado de múltiples programas de transcripción implementados en diferentes momentos. Aunque tienen la misma fuente de datos primaria, estos conjuntos de datos difieren con respecto a la cobertura de su estación, la integridad de la serie de tiempo y la calidad de la transcripción. Este estudio se realizó en el marco de los proyectos "Clima de datos y clima extremo para los Andes centrales" (DECADE) y "Servicios climáticos con énfasis en las arenas en apoyo a las decisiones" (CLIMANDES). Ambos proyectos apuntan a aumentar la calidad de los datos de observación en los Andes centrales, los conjuntos de datos disponibles fueron evaluados, fusionados, corregidos si es posible y actualizados. Una salida del proyecto DECADE será un conjunto de datos nuevo y mejorado para Bolivia y el sureste de Perú.

2.1. Redes de estaciones en Bolivia y Perú

Actualmente, la red de estaciones de SENAMHI Bolivia contiene alrededor de 255 estaciones tripuladas y 33 automáticas (F.Villalpando, 2016, comunicación personal). Tenga en cuenta que estos últimos no están incluidos en este trabajo. Para las series de tiempo digitalizadas que se utilizaron en este estudio, la longitud media de la suma de observaciones en cada estación de SENAMHI Bolivia es ligeramente inferior a 20 años para TX (Figura 1 (a)), TN y PRCP. Las distribuciones espaciales y temporales de las observaciones TX y TN son casi idénticas. Las distribuciones espaciales de la temperatura y las mediciones de PRCP son similares, pero el número total de series temporales de PRCP es aproximadamente el doble. Desde el año 2000, ha habido una disminución en el número de mediciones diarias de temperatura (Figura 1 (b)). Para PRCP, la disminución comienza ya a principios de la década de 1990, lo que resulta en aproximadamente la mitad de las observaciones en la última década en comparación con los máximos en las décadas de 1980 y 1990. Este patrón se observa en muchas redes de observación del clima en el mundo en desarrollo (Sene y Farquharson, 1998). Las razones de este desarrollo son múltiples e incluyen marcos institucionales inadecuados y la falta de apreciación del valor de los datos a largo plazo (Seneand Farquharson, 1998). Las mediciones digitalizadas más antiguas de SENAMHI Bolivia datan de 1917. Sin embargo, en el contexto de este estudio, se detectó un registro no digitalizado que data de 1892. Los datos climáticos históricos a menudo no se transcriben por completo (por ejemplo, Brunet y Jones, 2011, Trewin, 2013, Brönnimann, 2015). Actualmente, un programa de la Organización Internacional de Rescate de Datos Ambientales (IEDRO) está abordando este problema en Bolivia. (<http://iedro.org/activities/bolivia-data-rescue-program>).

Además de los datos de SENAMHI Bolivia, casi todos los registros de estaciones de AASANA se incluyeron en este estudio. Las observaciones se remontan a 1942. En contraste con la red SENAMHI Bolivia, la cantidad de mediciones diarias permanece casi constante desde la década de 1990 (Figura 1 (b)).



La mediana de la duración de la observación es de 40 años para temperatura y 52 años para precipitación. Se ha demostrado que recopilar y analizar datos de diferentes redes de la misma región es beneficioso para la precisión de los análisis meteorológicos (por ejemplo, Westerberget al., 2009). Las primeras mediciones en las series temporales peruanas disponibles para este estudio se tomaron en 1931. A diferencia de SENAMHI Bolivia, el número de mediciones diarias en SENAMHI Perú es más constante (Figura 1 (b)), y no se observa una disminución clara en las observaciones diarias en los últimos años. años, ya sea por temperatura o por PRCP. La duración media de observación es de 20 años para temperatura y 28 años para PRCP.

2.2. Prácticas de medición e instrumentación

La gran mayoría de las estaciones en los Andes centrales están tripuladas. Según Fiebrich y Crawford (2009), la automatización de una red da como resultado un claro aumento en la calidad de los datos. Las estaciones meteorológicas automáticas (AWS) han demostrado ser más precisas que las observaciones tripuladas para muchas variables (por ejemplo, temperatura), pero para otras (por ejemplo, PRCP), hay problemas de medición que deben abordarse (Ciach, 2003, Plummeret al., 2003 , Tokayet al., 2010). Sin embargo, AWS puede estar expuesto a perturbaciones, como defectos técnicos, robo o vandalismo. Dichas pérdidas de datos pueden hacer que las series de tiempo sean inutilizables para los análisis climatológicos (por ejemplo, Westerberget al., 2009). Además, las capacidades, el personal y el equipo deben ser adecuados para ejecutar la red (OMM, 2008). Por ejemplo, AWS requiere visitas a la estación al menos dos veces al año (Plummeret al., 2003). Este sería un gran desafío en la región andina central, donde el acceso a áreas remotas se ve obstaculizado estacionalmente y las inspecciones de la estación se realizan con muy poca frecuencia. Sin un estricto cumplimiento de los programas y programas de inspección y mantenimiento prescritos, la calidad de las observaciones de AWS puede deteriorarse dramáticamente (Plummeret al., 2003). Por lo tanto, una automatización repentina de las redes no solo puede causar una considerable falta de homogeneidad en los registros de medición (por ejemplo, Hubbard y Lin, 2006, WMO, 2008, Trewin, 2010), sino también una posible disminución de la calidad de las observaciones climáticas de los Andes centrales. Los observadores en Bolivia y Perú son en su mayoría laicos, que realizan las mediciones además de su ocupación principal. En algunas estaciones, los observadores oficiales también pueden ser reemplazados temporalmente por ayudantes (generalmente miembros de la familia), lo que resulta en cambios en la forma en que se recopilan los datos. Las mediciones de las estaciones SENAMHI Bolivia se toman a las 0800 LST (1200 UTC). Parcialmente, se toma una segunda medición diaria a las 1700 o 1800 LST. Los aeropuertos bolivianos miden cada hora durante los tiempos de operación, pero solo se transcriben los resúmenes diarios. En las estaciones de SENAMHI Perú, las mediciones se toman normalmente tres veces al día a las 0700 (1200 UTC), 1300 y 1900 LST. Los datos se transmiten en gran medida a SENAMHI en forma de documentos originales (en Bolivia cada dos meses), ya que es la práctica real o anterior en otras redes (por ejemplo, Kunkelet al., 2005). Actualmente, los observadores en algunas estaciones peruanas transmiten los datos inmediatamente a SENAMHI por teléfono móvil. En Bolivia, TX y TN se miden principalmente con seis termómetros. Estos termómetros se consideran robustos y fáciles de usar, pero fueron reemplazados por el trabajo de campo profesional en muchas redes hace mucho tiempo (Austin y McConnell, 1980). En general, solo los instrumentos de alta calidad deben utilizarse con fines meteorológicos (OMM, 2008). Sin embargo, como los instrumentos también deberían ser adecuados en cuanto a confiabilidad, durabilidad, simplicidad de diseño, los recursos para la capacitación de observadores y los costos (OMM, 2008), los termómetros de Six pueden ser la mejor opción para ciertas estaciones. En algunas estaciones bolivianas y en la mayoría de los peruanos, se utilizan termómetros de líquido en vidrio TX y TN. El PRCP se mide principalmente con pluviómetros Hellmann sin parabrisas. El uso de diferentes tipos de instrumentos para medir un solo parámetro puede causar problemas de continuidad de datos. (OMM, 2008).

Conclusiones:

Los problemas de calidad de los datos a menudo no se abordan adecuadamente y contribuyen a la incertidumbre en los resultados de los estudios climatológicos. Como muchos factores que afectan la calidad de los registros de observación (por ejemplo, una ubicación deficiente de la estación) son apenas detectables en los datos de medición, los metadatos son una fuente importante para evaluar la calidad de los datos. Sin embargo, en muchas redes de estaciones, los metadatos faltan y / o son de baja calidad. Para tales casos, este estudio sugiere un método diferente sobre cómo generar o mejorar metadatos. Se identificaron y atribuyeron los errores sistemáticos más comunes en las redes de estaciones predominantemente tripuladas de Bolivia y Perú. Se encuentran los mismos errores o similares en muchas otras redes de estaciones en todo el mundo, especialmente si las condiciones de observación son comparables. La mayoría de los problemas de calidad de los datos están relacionados con errores del observador. Por lo tanto, la forma más efectiva de prevenir la aparición de errores sistemáticos es mejorar el entrenamiento del observador y establecer un control de calidad casi en tiempo real que permita una intervención rápida en caso de que ocurra un error. Muchos enfoques de control de calidad tienen problemas para detectar problemas de calidad de datos que ocurren sistemáticamente. Estos métodos deberían mejorarse para abordar esta deficiencia. Sin embargo, la detección de tales problemas de calidad se dificulta por la alta variabilidad temporal de la ocurrencia de errores, la gran cantidad de especificidades de error y las interacciones de error. Como los problemas de calidad de los datos descritos en este artículo afectan la eficiencia de muchos algoritmos de CC, estos errores sistemáticos deben tratarse primero, y los parámetros de las pruebas estadísticas de CC deben adaptarse si es necesario. Se ha demostrado que la visualización de datos es una herramienta eficaz para identificar y atribuir problemas de calidad de datos. Proporcionar esta información permite a los usuarios de datos decidir para qué fines un registro de estación aún puede utilizarse, aunque se ve afectado por un error sistemático. En casos ideales, los datos defectuosos son incluso corregibles. Además, los segmentos de series temporales afectados se pueden eliminar de los registros de la estación y, para algunas aplicaciones específicas, se pueden aplicar métodos para ajustar los datos. La ganancia de series de tiempo utilizables logradas por este enfoque es particularmente importante en redes de estaciones dispersas. Todos los factores que afectan la calidad de los datos también pueden causar puntos de interrupción en las series temporales (OMM, 2008) y, por lo tanto, pueden dificultar la detección y corrección de otras inhomogeneidades (Domonkos, 2013). Esto es particularmente crítico en áreas con baja densidad de estaciones, como los Andes centrales, donde las bajas correlaciones de estaciones reducen en gran medida el rendimiento de los métodos de homogeneización de datos relativos, y donde los ajustes de datos potencialmente aumentan la falta de homogeneidad de los datos (Gubleret al., 2017; comunicación personal ) Debido a que la homogeneización de los medios mensuales o anuales puede no ser suficiente para las escalas de tiempo diarias (por ejemplo, Costa y Soares, 2009; Trewin, 2013; Brönnimann, 2015), la homogeneización diaria sería necesaria para estudiar los extremos que a menudo se ven más afectados por los problemas de calidad de los datos. . Sin embargo, esto requeriría correlaciones de estación aún mayores debido a la menor relación señal / ruido en escalas de tiempo más cortas. Además, los ajustes de datos estadísticos de los problemas comunes de calidad de datos presentados en este artículo a menudo no serían adecuados debido al carácter discreto de los errores. En resumen, la selección y el tratamiento de series de tiempo basadas en un control de calidad integral reducen el número de inhomogeneidades. Por lo tanto, esto puede contribuir a mejorar la homogeneidad de los datos en las redes andinas centrales y comparables más que los métodos de homogeneización de datos estadísticos. Sin embargo, como siguiente paso, la aplicación cuidadosa de la homogeneización de datos estadísticos puede aumentar aún más la homogeneidad de los datos. Los hallazgos sobre los problemas de calidad de datos que se presentan sistemáticamente en este estudio pueden ser útiles para la creación de conjuntos de datos de referencia realistas para el control de calidad y los algoritmos de homogeneización. , no. IZ01Z0\_147320, que está financiado por el Programa Suizo de Investigación sobre Asuntos Globales para el Desarrollo (r4d), y el proyecto "Servicios CLIMáticos con énfasis en Los Andes en Apoyo a las decisiones" (CLIMAN-DES), no. 7F-08453.01, financiado por la Agencia Suiza para el Desarrollo y la Cooperación (COSUDE) y coordinado por la Organización Meteorológica Mundial (OMM). Agradecemos además a Jenny Chavez, Roberto Catacora y Fernando Figeroa de AASANA por su gran apoyo, Andrew Rhines por la útil conversación sobre la decodificación de las precisiones de medición, y Paulo Vitor DaCosta Pereira y Yuri Brugnara por la información e intercambio sobre problemas de calidad de datos en Redes brasileñas e italianas, respectivamente. SB reconoce el apoyo del proyecto FP7 ERA-CLIM2.