



Servicio
Meteorológico
Nacional

Plan transversal de verificación de pronósticos del SMN

Nota Técnica SMN 2021-84

Ramón de Elía¹

¹ *Dirección Nacional de Ciencia e Innovación en Productos y Servicios, SMN*

Enero 2021



Ministerio de Defensa
Presidencia de la Nación

Información sobre Copyright

Este reporte ha sido producido por empleados del Servicio Meteorológico Nacional con el fin de documentar sus actividades de investigación y desarrollo. El presente trabajo ha tenido cierto nivel de revisión por otros miembros de la institución, pero ninguno de los resultados o juicios expresados aquí presuponen un aval implícito o explícito del Servicio Meteorológico Nacional.

La información aquí presentada puede ser reproducida a condición que la fuente sea adecuadamente citada.

Resumen

Las actividades de verificación de pronósticos dentro del SMN han sido realizadas hasta recientemente de forma independiente en cada una de las divisiones responsables de pronósticos. El Plan transversal presentado aquí apunta a generar elementos comunes que ayuden en el intercambio de ideas, de metodologías y que permitan la comparación entre los diferentes pronósticos cuando esto es factible. Esto último es de particular relevancia en lo que respecta a pronósticos que son parte de la misma cadena de valor, por ejemplo los de temperatura máxima a 24 hs en una dada ciudad, realizados tanto por los modelos globales, como por los regionales, como por el personal a cargo de emitir los pronósticos oficiales.

Abstract

Forecast verification within the Weather office of Argentina (SMN) was carried out till now independently by each division responsible for the production of public forecasts. The verification plan presented here aims at concentrating on the common points, ideas and methodologies that share the different verification practices, as well as allowing for intercomparison when this is possible and desirable. This intercomparison is especially important with respect to forecast that are part of the same value chain, as for example the maximum temperature in a 24h-forecast at a given city, produced by global and regional models, as well as by the personnel in charge of subjective forecast products.

Palabras clave: pronósticos, verificación, plan

Citar como:

de Elía R., 2021: Plan transversal de verificación de pronósticos del SMN. Nota Técnica SMN 2021-84.

1. INTRODUCCION

Este plan tiene como objetivo establecer criterios y metodologías comunes entre diferentes Direcciones del SMN en lo que respecta a la verificación de pronósticos. Si bien cada tipo de pronóstico tiene sus necesidades, hay varios elementos que son comunes o, lo que es más relevante, están interconectados. Por ejemplo, la verificación de pronósticos numéricos tiene que relacionarse con las necesidades de sus usuarios, como es por ejemplo el caso de la aeronáutica que tiene sus necesidades específicas impuestas por la OACI(ver OACI 2007).

Las variables y las metodologías que se podrían utilizar para verificar los pronósticos del SMN son ilimitadas, y aumentan en número con el correr de los años. Para evitar naufragar en cifras y que cada institución o grupo derive en la adopción de un conjunto particular de índices, la OMM ha instado desde hace varias décadas a la uniformización de los métodos, y esto se formalizó con la creación en 2003 del *WWRP/WGNE Joint Working Group on Forecast Verification Research*.

Como toda uniformización tiene sus riesgos, no se busca que todos los grupos estén obligados a aplicar las mismas recetas sino a que todos produzcan mínimamente un grupo de índices que permita comparar resultados y hacer discusiones inteligibles. Es importante notar la importancia de que estas comparaciones sean factibles no sólo entre instituciones, sino también entre épocas. Es decir que esto nos obliga a mantener una coherencia temporal en nuestra manera de trabajar si queremos que nuestros resultados de hoy sean comparables con los de las próximas décadas.

De este tipo de recomendaciones existen varias y surgen algunas variaciones con el tiempo, con el agregado de nuevas métricas a medida que estas se vuelven más populares o que nuevos tipos de verificaciones se hacen necesarias. Con el objetivo de establecer estas normas dentro del SMN se propone aquí centrarse sobre un texto particular de recomendaciones de verificación. El texto elegido es *Recommendations on the verification of local weather forecasts* de Pertti Nurmi (2003), que tiene varias características que lo hacen apropiado: La primera es que son recomendaciones producidas por el *European Center for medium-Range Forecasting* (ECMWF), un organismo internacional que reúne países de muy diferente capacidad en lo que refiere a meteorología, desde el Reino Unido pasando por Portugal y Turquía hasta Grecia. Es decir, no propone prácticas irrealizables en un país como Argentina. La segunda es que el texto fue publicado hace casi dos décadas pero sigue vigente, y es citado con frecuencia. La tercera es que propone una visión bastante completa en lo que a tipos de pronósticos se refiere, a pesar de provenir de un centro dedicado mayormente al pronóstico numérico al mediano plazo. La cuarta es que se centra en la verificación de pronósticos locales –los que afectan a nuestros usuarios–, y no en mediciones que, siendo muy útiles para comparar performance de modelos, son necesariamente abstractas. En general se postula la verificación como una herramienta de doble uso: por un lado para conocer los desvíos de los modelos como insumo para disparar la ciencia que los hace evolucionar, y por otro como instrumento de comparación y de rendición de cuentas de las instituciones. Este documento se centra sobre el segundo, pero toca ocasionalmente temas relacionados al primero.

Es importante reiterar y destacar que con el objetivo de realizar un plan transversal que involucre a todos los pronósticos a ser realizados en el SMN, este documento se centra sobre pronósticos locales, dejando mayormente de lado otros elementos de gran importancia meteorológica como los patrones espaciales y variables en altura. Por la misma razón, la verificación de pronósticos por ensambles no juega un rol destacado. Para esto último se sugiere al lector ir a Wilks (2006).

2. LINEAMIENTOS GENERALES

2.1 Cadena de pronósticos

Entre los principios que rigen este grupo de recomendaciones, está naturalmente la necesidad de uniformizar criterios para establecer comparaciones y diálogo. Rige además la voluntad de detectar el “valores agregado” de cada elemento de la cadena de pronósticos, desde los modelos globales a los regionales, a los modificados por postprocesamiento (como las calibraciones), hasta los realizados por los pronosticadores tomando a estos como insumos. Es decir, no solo se piensa la verificación como “producto final” vs “observaciones”, sino en cada instancia de la cadena de producción.

2.2 Consideración de la resolución

Es importante recordar que la comparación de pronósticos de diferente resolución redunda en resultados que tienden a desfavorecer los pronósticos de más alta resolución. Este llamado efecto de la “doble penalización” (en inglés *double penalty*) es más agudo con variables de alta variabilidad en las altas frecuencias espacio temporales (vorticidad, precipitación de verano, etc.; ver por ejemplo Zingerle y Nurmi 2008, y explicación asociada en su Figura 6). Esto puede ser entendido como el efecto de pronosticar en alta resolución un evento que no fue pronosticado por un modelo de baja resolución --lo cual es un éxito--, pero si este aparece ubicado en el tiempo o lugar equivocado, se paga no solo por su ausencia en el lugar que debería estar, sino por su presencia en donde no debería. De allí lo de penalización doble.

2.3 Las observaciones

Desde las últimas décadas han comenzado a aparecer numerosas fuentes de observaciones que se suman a las clásicas de las redes de los servicios meteorológicos. Si bien estas nuevas fuentes como los sensores remotos son de compleja utilización es recomendada una evolución hacia su uso como elementos de verificación. Durante el proceso de escritura de Nurmi (2003) no existían las redes sociales, pero estas sí son un desafío que al SMN le interesa particularmente, y por lo tanto podrían considerarse como una de las exploraciones que pueden ser de particular interés para nuestra institución.

2.4 Representatividad de las observaciones

El llamado “problema de la representatividad” de las observaciones describe el hecho de que cada instrumento mide una parcela particular de la atmósfera cuya forma no es siempre fácil de definir (ver Horel 2008, López y otros 2011). La medición instantánea de un termómetro tiene una alta correspondencia con la temperatura del terreno circundante, pero el tamaño de este territorio depende de las circunstancias. Un caso más complicado es el de la precipitación, que tiene una compleja estructura espacio temporal. La medición de un pluviómetro se ocupa exactamente de la lluvia caída en un círculo de unos centímetros de diámetro, que poco puede tener que ver con lo acumulado a unos kilómetros de distancia. Naturalmente esto tiene un efecto a la hora de verificar pronósticos, donde se hace la suposición que se compara elementos de igual representatividad (por ejemplo precipitación acumulada en 6 hs en un pluviómetro con un punto de grilla del modelo). Es importante notar que este problema se agrava al no haber claridad en el mundo de la modelización sobre la representatividad espacial de una variable en un punto de grilla (ver por ejemplo Pielke 2001).

Muchos de estos desfasajes se pueden corregir con postprocesamiento de datos, pero en general deben aceptarse como una fuente alternativa de error, que no corresponde ni a las observaciones ni a los pronósticos.

2.5 Agregación y desagregación

Dados los problemas de representatividad y de falta de datos, muchas veces es necesario amalgamar observaciones, o puntos de grilla, o pasos de tiempos, o fenómenos para dar una más apropiada representación, sea de los pronósticos o de la realidad observada. Por ejemplo, durante el invierno las tormentas fuertes son poco frecuentes sobre CABA, y si se quiere tener una estadística más robusta -- mayor número de casos-- para verificar tal fenómeno sería preferible extender la región de interés a un área más grande que la rodea.

De manera inversa, en muchas ocasiones lo recomendable es desagregar, sobretudo si la muestra es lo suficientemente grande como para permitirlo. Estas desagregaciones suelen proveer información adicional. Por ejemplo, si los números de casos disponibles lo permiten, no es solo interesante saber si uno tiene habilidad (skill) con las temperaturas máximas durante el verano, sino también, por ejemplo en casos particulares como durante los días nublados.

Cuando la cantidad de casos lo permite, desagregar ayuda a hilar más fino sobre las capacidades de los pronósticos. Las siguientes son las desagregaciones/agregaciones más frecuentes:

En el tiempo: anual, estacional, mensual, ciclo diurno, horizonte del pronóstico

En el espacio: efecto topográfico, contraste tierra agua, regiones de diferente tamaño.

Umbrales: valores medios, extremos

2.6 Pronóstico de referencia

Todo pronóstico recibe su certificado de utilidad al ser comparado positivamente con otro. Esta comparación está en el ADN del progreso de los pronósticos meteorológicos. De los pronósticos imaginables hay dos muy simples contra los que todo pronóstico no puede evitar ser comparado: la climatología y la persistencia. Cuando es posible, todo índice de verificación (score) debe estar acompañado por su ventaja sobre un pronóstico de referencia. Mínimamente, los dos mencionados; de máxima todos los que son parte de la cadena de valor, mostrando, por ejemplo, que un pronóstico subjetivo agrega algo a uno automático.

2.7 Robustez estadística

Todo índice de verificación tiene que poder ser interpretado como correspondiente al sistema de pronóstico y no como producto del azar. Para ello es importante presentar estimaciones de su robustez estadística (por ejemplo, intervalo de confianza), y no presentar resultados cuyos robustez sea demasiado baja. Es mejor esperar a que la muestra crezca o realizar una agregación de datos en tiempo, umbrales o espacio (como discutido en sección 2.5). Una breve discusión sobre este punto se puede encontrar en el Apéndice A.

2.8 Verificación en un entorno operacional

En un entorno operacional es importante tener los índices de verificaciones automatizados y que sean presentados periódicamente en informes. Un sistema de pronóstico necesita una gran serie de datos para mostrar su efectividad, por lo que la periodicidad de los reportes debe ser evaluada con precaución. Algunos centros presentan estadísticas diarias o mensuales (ver por ejemplo <https://sats.nws.noaa.gov/~verification/ndfd/>), pero es clave en estos casos remarcar su valor circunstancial.

Por ejemplo, entre mes y mes un sistema de pronósticos puede responder de manera dispar. Esta disparidad puede tener diferentes fuentes: si hay intervención humana algún cambio sustancial en el personal, algún cambio en las herramientas utilizadas, o simplemente el azar. En una muestra pequeña este último efecto es tan importante que oscurece casi cualquier otra fuente. El riesgo de ignorar este punto puede hacer atribuir a causas imaginarias las diferencias en verificación observadas.

2.9 Comunicabilidad de los índices

Más allá de la necesidad de contar con los índices que se discuten a continuación que pueden usarse con diversos fines, es política del SMN disponer siempre de índices que sean de fácil comunicación. Por ejemplo, no es lo mismo comunicar una tasa de éxito --cantidad de casos acertados para un dado número de casos--, que un error cuadrático medio. En muchos casos, cuando no exista un índice evidente que facilite la comunicación, se debe alterar la variable de interés hasta encontrar el apropiado. Por ejemplo, en el caso de un pronóstico probabilístico de precipitación puede tomarse un umbral de probabilidad y transformarse en un pronóstico determinístico. Como dice en las *Guidelines on Performance Assessment of Public Weather Services* (WMO 2000, pág. 2), es muy probable que los usuarios nunca hayan oído hablar del Brier score, y también es probable que no sean capaces de transmitirlo a otros.

Es importante notar que índices como la tasa de éxito, si bien son simples de comprender, son por otro lado muy sensibles a los bias sistemáticos de los pronósticos, por lo cual es muy importante tener esto en cuenta para decidir si es conveniente su adopción.

2.10 Simplicidad

Incluso cuando los scores e índices son para consumo interno del SMN es importante utilizar criterios simples. En las *Guidelines on Performance Assessment of Public Weather Services* (WMO 2000, página 7), se puede leer: “Embarcarse en un programa de verificación es un gran desafío para un SMHN con poca experiencia en la materia. Es mejor utilizar medidas simples y fáciles de entender que implementar metodologías complejas. Es mejor concentrarse solo en algunos lugares claves que tratar de verificar muchas variables pronosticadas en muchos lugares. Manteniendo un número bajo de verificaciones evita sumergirse en números que nunca son analizados y que resultan una pérdida de tiempo producir”.

3. VARIABLES CONTINUAS

Ya que la metodología de verificación difiere de manera importante entre variables continuas y discretas se propone discutir las separadamente. Empezamos aquí por las continuas.

3.1 Variables elegidas

Las variables más habituales que se verifican son las siguientes (notar que son variables en superficie).

Temperatura en horas fijas del día, T_{min}, T_{max}, y sus promedios espacio temporales.

Dirección e intensidad del viento en horas fijas del día, y sus promedios espacio temporales.

Precipitación acumulada integrada en el tiempo (por ejemplo 6, 12, 24 hs)

Nubosidad en tiempos fijos (notar que a pesar de ser continua, en general se la trabaja de manera discreta, por lo cual se adapta mejor a lo discutido en la sección que sigue)

Estas variables mencionadas están íntimamente relacionadas con las actividades humanas. Cabe notar la ausencia de variables estrechamente ligadas a la humedad atmosférica --como las temperatura de rocío (T_d) o cualquiera de sus variables asociadas--, que es generalmente ignorada por los pronósticos al público (la humedad es una variable un poco abstracta), salvo en casos donde esta afecta claramente la actividad humana, como por ejemplo en días de calor de alta humedad. Es importante también notar que el pronóstico de temperatura mínima puede ser entendido como un sustituto (*proxy*) del pronóstico de variables relacionadas con la humedad.

3.2 Diagramas de carácter exploratorio

La utilización de índices de verificación tiene la ventaja de reducir a muy pocos números lo esencial de un trabajo de verificación. Esto naturalmente facilita la creación de “mensajes” a la hora de comunicar los mismos. Como todas las estadísticas, el poder de síntesis de los índices es también un ejercicio en medias verdades --sobre todo cuando se utilizan pocos de estos índices. En el extremo contrario de estos índices se hallan los diagramas, que carecen de poder de síntesis pero suelen ser muy elocuentes a la hora de explicar el origen de los resultados.

Entre los que son de utilidad y que se discuten más abajo tenemos:

Scatterplots: en particular **observaciones vs. pronósticos**, **observaciones vs. (pronósticos - observaciones)**, y **pronósticos vs. (pronósticos - observaciones)**, permiten detectar desvío de los pronósticos a diferentes umbrales con un simple golpe de vista.

Series de tiempo: Ideal para detectar existencia de pronósticos anómalos, periodicidades y tendencias.

3.3 Índices

Existen muchos índices que tienden a visibilizar diferentes aspectos de la relación pronóstico-observación. Los más utilizados son

Error medio (bias)

Error medio absoluto

Error cuadrático medio

Más detalle sobre estos y otros scores se pueden encontrar en Nurmi (2003), y en el sitio web <https://www.cawcr.gov.au/projects/verification/> que contiene una importante lista de referencias.

Aunque no mencionado en Nurmi (2003), resulta de interés en algunos casos descomponer el error cuadrático medio, en bias, varianza y correlación. Esta descomposición además de agregar información tiene la ventaja de poder ser graficada de manera muy informativa, ideal para agrupar por localidad (ver Taylor 2005).

4. VARIABLES CATEGÓRICAS

Las variables categóricas juegan un rol clave en meteorología, en particular porque toda variable continua puede además pensarse como dividida en categorías. Esto tiene varias razones para hacerse, pero en especial se lleva a cabo como una manera de amalgamar datos en grupos (intervalos) evitando dispersar datos en todo el eje continuo. Esto se relaciona con la búsqueda de robustez de resultados, como fue

comentado antes. En lo que sigue se muestran las variables elegidas cuya división de categorías es tanto natural como un recurso para simplificar su análisis.

4.1 Variables elegidas

Lluvia (vs. no lluvia), con varios umbrales

Nieve caída (vs. no nieve), con varios umbrales

Vientos fuertes (o ausencia de estos), con varios umbrales

Helada (vs no helada)

Niebla (vs. no niebla)

Si bien Nurmi (2003) no introduce la nubosidad en este grupo, aquí se la considera como parte del mismo.

4.2 Tabla de contingencia

La tabla de contingencia es la herramienta por excelencia para estudiar pronósticos categóricos. Aquí abajo se muestra el muy popular caso de la tabla para pronósticos dicotómicos (solo dos categorías).

		Observado		
		SI	NO	Totales
Pronóstico	SI	N° de Aciertos	N° Falsas alarmas	Total pronosticados SI
	NO	N° de Sorpresas	N° Aciertos negativos	Total pronosticados NO
Totales		Total observados	Total no observados	Total muestra, pares pronóstico/observación

Los cuatro resultados se interpretan de la forma siguiente:

Acierto: se pronostica un evento y este ocurre. Este evento puede ser tanto un auténtico evento binario (llueve, no llueve), como la binarización de un evento continuo. Por ejemplo, se puede definir como “aciertos” aquellos eventos donde la temperatura pronosticada y observada son mayores a un cierto umbral.

Falsa Alarma: se pronostica un evento y este no ocurre

Sorpresa: no se pronostica evento --o se pronostica que no ocurrirá--, pero este ocurre

Acierto negativo: no se pronostica evento --o se pronostica que no ocurrirá--, y este no ocurre.

Es importante tener en cuenta que el registro de eventos en una red con estaciones espaciadas y en regiones de baja densidad demográfica, tiende a subestimar la cantidad de eventos reales. Esto naturalmente afecta a las categorías “Falsas alarmas” y “Aciertos negativos”, que en realidad podrían llamarse “Falsa alarma o acierto no registrado” y “Aciertos negativos o sorpresa no registrada”. La realidad de cada caso debe ser considerada con atención por quién realiza la verificación.

4.3 Índices de la tabla de contingencia

Si bien la tabla de contingencia tiene un aspecto muy simple --son en definitiva cuatro cifras que definen las características del pronóstico dicotómico-- esta apariencia es engañosa: son decenas los índices que pueden postularse a partir de estos números, y todos ellos responden a preguntas o conceptos diferentes (ver por ejemplo <https://www.cawcr.gov.au/projects/verification/>). Es difícil luego de varios años trabajando en el tema no sentir la tentación de inventar su propio índice o mostrar una predilección subjetiva por uno en particular. Para evitar esos desvíos se propone centrarse en los mencionados abajo.

Bias en frecuencia
Proporción correcta
Probabilidad de detección
Proporción Falsas alarmas
Probabilidad de falsas detecciones
Hanssen-Kuipers Skill Score (KSS)
Threat score
Equitable Threat score
Heidke Skill score (HSS)

Todos estos scores están basados en las cuatro cifras mencionadas en la sección anterior, pero no todos utilizan las cuatro (ver Apéndice B para el detalle de los más populares). Esto es interesante porque no todos los pronósticos llenan todos los casos, por diferentes razones. Por ejemplo, la casilla de “Aciertos negativos” no existe en el pronóstico de Avisos a corto plazo (ACPs) del SMN (no se emiten pronósticos diciendo “no hay ACP”, por lo cual no es obvio cómo obtener esa categoría (ver Apéndice C para más detalles sobre este punto, y San Martino y otros 2019 para más información sobre los ACPs).

4.4 Pronósticos de más de dos categorías

Este caso nos conduce a construir tablas de contingencias más grandes. En el SMN existen al menos los siguientes casos de pronósticos de más de dos categorías.

Niebla (nada, neblina, niebla)
Nubosidad (despejado, algo nublado, parcialmente nublado, mayormente nublado, nublado)
Intensidad de precipitación: 4 categorías para lluvias persistentes: nada, llovizna, lluvias, lluvias fuertes; y 4 para lluvias intermitentes: nada, chaparrones, tormentas aisladas, tormentas fuertes)
Tormentas (3 categorías, nada, fuerte, severas)

De los índices mencionados para el caso de tablas de contingencia de 2x2, KSS y HSS pueden ser extendidos a tablas de cualquier tamaño y son por eso recomendados en estos casos (ver Wilks 2006).

5. PRONÓSTICOS PROBABILÍSTICOS

Los pronósticos probabilísticos tienen por objetivo incluir información sobre la incerteza asociada al pronóstico y, para algunas variables, se los utiliza desde hace años en muchos países del mundo (ver por ejemplo Murphy 1998). Suele utilizarse en variables u horizontes donde el nivel de incertidumbre es alto, por ejemplo la precipitación a cualquier plazo o la temperatura a escala trimestral.

Existen muchas metodologías para verificar los pronósticos probabilísticos, la más utilizada y mejor fundamentada es la asociada a los desarrollos de Allan Murphy, que comienza en Murphy (1973) y produce una larga serie de exitosos papers que hicieron escuela. De allí surge el muy utilizado reliability diagram, y la descomposición del Brier score en atributos esenciales. A partir del Brier score y de su comparación con los scores de pronósticos de referencia (la climatología, la persistencia, modelo global piloto, etc) se define el muy utilizado Brier skill score (BSS). Para una breve discusión ir a <https://www.cawcr.gov.au/projects/verification/>.

Una extensión del Brier score a varias categorías es el llamado Ranked Probability Score (RPS) y su versión "score" el Ranked Probability Skill Score (RPSS). Este es particularmente útil en principio cuando se trata de pronósticos probabilísticos del tipo realizado por categorías (como en el caso pronóstico estacional), pero su uso no es siempre recomendado (ver Mason 2018).

También suele utilizarse el diagrama ROC (Relative Operating Characteristic), que tiene la ventaja de estar compuesto de elementos de alto nivel de comunicabilidad, ya que se relaciona estrechamente con el nivel de aciertos y de falsas alarmas descrito en la sección 4.3.

Es importante remarcar que el pronóstico probabilístico es una de las posibilidades que surgen naturalmente del uso de ensambles, pero la verificación específica de los ensambles merece una discusión que excede a este documento (ver Wilks 2006).

6. PLANES POR AREA DENTRO DEL SMN

Cada área del SMN que produce pronósticos de algún tipo deberá construir sus sistema de verificación utilizando este documento como base. Lo más importante de este documento son sus ideas generales que pueden aplicarse a todos los pronósticos. Los puntos específicos naturalmente no cubren todas las posibilidades, por lo cual en cada área habrá una componente de desvío sobre las recomendaciones de este documento. Estos desvíos, no obstante, deben acercarse lo más posible al espíritu de las recomendaciones generales.

6.1 Pronósticos para los próximos 7 días (DNPSS/DPTA/CPR)

Las recomendaciones aquí discutidas se adaptan bien a las actividades de pronósticos convencionales de esta Coordinación. Será necesario no obstante establecer cuáles de las variables aquí discutidas son factibles de verificación y bajo qué condiciones.

Dentro del mundo de los pronosticadores operativos --y el SMN no es excepción-- es usual utilizar los pronósticos de temperatura donde el "acierto" se define como un intervalo de $\pm 2^{\circ}\text{C}$ alrededor del valor observado (ver por ejemplo Cattani y otros 2015, o página web del servicio meteorológico de Nueva Zelanda y de Australia; el reporte WMO (2000) también indica que es difícil escapar a este tipo de índice por su alta claridad para el público y lo pone como opción, ver páginas 9 y 10). Esta metodología es muy similar a la impuesta por la OACI para valores aceptables de predicción de temperatura (más del 70% de acierto en un intervalo de $\pm 1^{\circ}\text{C}$, ver sección 6.5).

Recomendación preliminar: *Productos de verificación* (mensual, estacional, anual, multianual a diferentes horizontes (en general 24hs, 48hs, 72hs, CABA hasta 168hs), por ciudad, por dominio, todo el país)

- . Tablas de contingencias precipitación (dicotómica) con scores (algunos de los mencionados en 4.3)
- . Reliability diagrams de probabilidad precipitación (con inclusión de distribución de pronósticos), Brier Skill Score
- . Scatterplot $(T_{min}-T_{min}^o)/T_{min}^o, (T_{max}-T_{max}^o)/T_{max}^o$, bias y error cuadrático medio
- . Probabilidad de detección de T_{min} , T_{max} , (dicotómicas) con índices (cuando la T pronosticada está dentro de un intervalo de $\pm 2^{\circ}\text{C}$ del observado).
- . Series temporales
- . Mapas de índices
- . Tabla de contingencia 5x5 de intensidad de viento
- . Intervalos de confianza

6.2 Pronóstico numérico con WRF y GFS (DNCIPS/DMSR)

Las recomendaciones aquí discutidas se adaptan sólo parcialmente a las actividades de verificación habituales con pronósticos numéricos, ya que estas tienen un interés centrado en el desarrollo o conocimiento de los modelos. No obstante, es importante que los usuarios internos del SMN tengan las verificaciones de pronósticos que funcionan como insumo para realizar sus propias verificaciones. En particular, debe responderse a las necesidades de DNPSS/DPTA (punto 6.1 y 6.7) y (DNPSS/CPI) (punto 6.5).

Hay tres temas que este documento deja esencialmente de lado: los pronósticos de patrones espaciales, la verificación en altura, y la temática general de los ensambles.

La precipitación no solo es de gran importancia socio-económica sino que además posee una gran complejidad. Por esta razón la precipitación --en particular cuando nos interesan escalas cada vez más pequeñas y pronósticos cuantitativos-- escapa a la verificación por métodos simples. Un documento que se centra sobre este desafío es *Recommendations for the Verification and Intercomparison of QPFs and PQPFs from Operational NWP Models* (WMO 2004).

En variables de altura tiende a verificarse geopotencial, temperatura, humedad relativa, y velocidad del viento a niveles específicos, así como también presión a nivel del mar (ver por ejemplo NWS <https://sats.nws.noaa.gov/~verification/ndfd/> y <https://sats.nws.noaa.gov/~verification/mos/>, CPTEC <http://avaliacaodemodelos.cptec.inpe.br/>).

La WMO emitió recomendaciones para la verificación de modelos determinísticos regionales en su reporte de 2010 donde diferentes niveles en altura son recomendados a verificar, dependiendo de la variable de interés (ver WMO 2010a, apéndice 2.2.34).

En general estas verificaciones se hacen en dos modos: los llamados GRID-to-OBS (pronóstico en retícula vs. estaciones o radiosondas) y GRID-to-GRID (pronóstico en retícula vs. análisis). Aquí nos concentramos sobre el primer tipo. Ver <https://www.emc.ncep.noaa.gov/users/verification/global/gfs/ops/>.

A este respecto abajo se muestran resultados de una encuesta en cuatro centros meteorológicos sobre las herramientas utilizadas en verificación. Se puede notar diferencias, pero también puntos comunes.

Centro	Herramientas
ECMWF	<u>BSS</u> , <u>reliability diagram</u> , <u>ROC</u> , ROC area, econ. value, spread/skill diagram
NCEP	<u>RMSE</u> and AC of ensemble mean, <u>BSS</u> , <u>ROC</u> area, <u>rank histogram</u> , RPSS, econ. value
Met Office	<u>BSS</u> , <u>reliability diagram</u> , <u>ROC</u> , <u>rank histogram</u>
BMRC	<u>RMSE</u> ensemble means, <u>BSS</u> , <u>reliability diagram</u> , <u>ROC</u> , <u>rank histogram</u> , RPSS, econ. value

Sacado de Ebert (2005) IGGE workshop, 1-3 March 2005, ECMWF. <https://www.ecmwf.int/sites/default/files/elibrary/2005/15865-verification-ensembles.pdf> (el subrayado muestra las herramientas mencionadas en este texto)

Recomendación preliminar: *Productos de verificación* (mensual, estacional, anual, multianual a diferentes horizontes, por ciudad, todo el país)

- . Tablas de contingencias precipitación (dicotómica) para distintos umbrales, scores (mencionados en 4.3). Distribución espacial en mapas
- . Reliability diagrams de probabilidad precipitación (con inclusión de bias)
- . Bias y RMSE de Temperatura. Distribución espacial en mapas
- . Probabilidad de detección de Tmin, Tmax, (dicotómicas) con scores (cuando la T pronosticada está dentro de un intervalo de $\pm 2^{\circ}\text{C}$).
- . Bias y RMSE de Temperatura de rocío. Distribución espacial en mapas
- . Bias RMSE magnitud del viento. Distribución espacial en mapas
- . Tabla de contingencia 5x5 de intensidad de viento
- . Scatterplot Tmin, Tmax, Td
- . Intervalos de confianza
- . Comparacion con modelo piloto

6.3 Pronóstico a muy corto plazo (DNPSS/DPTA/CPI)

Este texto tiene aplicación limitada para la verificación de ACPs, pero sí establece una buena base formal que sirve de guía. Es indudable que llegar a verificaciones de ACPs va a implicar desarrollos internos ya que no hay mucho desarrollado a nivel internacional y menos aun recomendaciones centralizadas de instituciones. El SMN cuenta con recursos humanos formados para encarar este trabajo necesario, acelerado por la utilización de redes sociales como instrumento de verificación.

Recomendaciones específicas para este tipo de pronóstico se encuentran en el capítulo 3 de *Guidelines for Nowcasting Techniques* (WMO, 2017) y también en Stephenson y otros (2010). Quizás debiera encararse este trabajo en conjunto con la verificación de Alertas (sección 6.8) dados los varios puntos comunes. Una importante característica de estos pronósticos es la dificultad de estimar los “aciertos negativos” en la tabla de contingencia, lo cual tiene importantes implicancias en las herramientas disponibles para verificación. Para mayor detalle ver Apéndice C.

Recomendación preliminar: *Productos de verificación* (mensual, estacional, anual, multianual centrado en grandes ciudades)

- . Tabla de contingencia de ACP para toda la República Argentina
- . Tabla de contingencia de ACP para cinco centros urbanos cubriendo estos de manera parcial o total (Gran Buenos Aires - Gran La Plata; Córdoba Capital; Santa Fé - Paraná; Mar del Plata y Rosario)
- . Utilización de scores que no contabilicen la categoría “aciertos negativos”, no recomendable en estos casos: Bias en la frecuencia, Proporción correcta (modificado, similar a “success ratio”, no recomendado), Probabilidad de detección, Proporción de Falsas alarmas, Threat score, (ver sección 4.3 y <https://www.cawcr.gov.au/projects/verification/>).
- . Diagrama de Performance (ver Apéndice C)
- . Categorización de observaciones: en METAR, SPECI, SYNOP, EMAs, reportes periodísticos, redes sociales.
- . Intervalo de confianza.
- . Estratificación en fenómenos reportados: Tormenas fuertes, lluvias intensas, rafagas, granizo
- . contabilidad de ACPs emitidos con errores técnicos

6.4 Pronósticos estacional (DNCIPS/DCMC)

Poco de lo aquí discutido se adapta las necesidades de CRC-SAS, y por esta razón en este caso vale la pena regirse por las sugerencias de Mason (2018). Sería conveniente para profundizar en el conocimiento de los índices, no obstante, que se utilice lo más posible lo recomendado en este texto. En la actualidad se utilizan scores menos convencionales, como el Ignorance Score (IS) y el Interest rate score (IR). El Climate Prediction Center utiliza el Heidke skill score, que sí aparece en la lista de arriba y es sugerido también por Mason (ver

https://www.cpc.ncep.noaa.gov/products/predictions/long_range/tools/briefing/seas_veri.grid.php).

Recomendación preliminar: *Productos de verificación* (estacional, anual, multianual centrado en regiones, por categoría, por variable) --inspirado en Mason (2018), adaptado a la lógica de este texto.

- . Reliability diagrams
- . Descomposición del Brier Score
- . Ignorance score
- . Combinación de diferentes localidades para ganar robustez
- . Heidke (Hit) score y skill score (para las tres categorías)
- . ROC por categoría, por variable
- . Para mapas individuales Heidke score
- . Intervalos de confianza

6.5 Pronóstico aeronáutico (DMA)

La OACI tiene sus propios criterios para las verificación de pronósticos que si bien no se oponen a los principios aquí discutidos, son más específicos. Esto no es solo el caso con respecto a la metodología sino también en lo que respecta a los *valores* de los índices recomendados. Es decir, hay claras demandas sobre la calidad de los pronósticos. Esto se puede ver en el Adjunto B del Anexo 3 de la OACI (2007). Un trabajo interesante sobre medidas de verificación en la práctica es Mahringer (2008). La tabla de abajo muestra las recomendaciones de la OACI con respecto a los TAF (Terminal Area Forecast). Notar que el índice sugerido es la tasa de éxito.

Demanda OACI: <i>Productos de verificación</i> TAF (propiedades de tasas de éxito, aeropuertos elegidos)	
Dirección del viento	$\pm 20^\circ$ (80% de los casos)
Velocidad del viento	± 10 km/h (5 kt) (80% de los casos)
Visibilidad	± 200 m hasta 800 m; $\pm 30\%$ entre 800 m y 10 km (80% de los casos)
Precipitación	Ocurrencia o no (80% de los casos)
Cantidad de nubes	Una categoría por debajo de 450 m (1 500 ft) Ocurrencia o no de BKN u OVC (70% de los casos)
Altura de las nubes	± 30 m (100 ft) hasta 300 m (1 000 ft) $\pm 30\%$ entre 300 m (1 000 ft) y 3 000 m (10 000 ft) (70% de los casos)
Temperatura	$\pm 1^\circ\text{C}$ (70% de los casos)

6.6 Pronóstico semanal por análogos (DNCIPS/DCMC)

Dado el rol que juegan los datos de las estaciones para “calibrar” los pronósticos por análogos, podemos considerarlos como pronósticos locales. Siendo así, las recomendaciones aquí discutidas se aplican perfectamente, con la salvedad de que son variables menos convencionales: temperatura media semanal, y precipitación acumulada en una semana. Ambos son pronósticos de variables continuas.

Es importante en este tipo de pronósticos que implican las primeras dos semanas, la selección apropiada de índices. La extensión del período implica baja predictibilidad y por lo tanto uno estaría tentado a utilizar índices más deterministas al inicio, mientras que hacia el final del periodo podrían asemejarse más a aquellos utilizados en pronósticos estacionales. Es por ello clave seguir de cerca el avance de la ciencia en este tema (ver Ou y otros 2016). Es importante también coordinar con los pronósticos semanales que utilizan el GEFS como fuente (sección 6.7)

Recomendación preliminar: *Productos de verificación* (estacional, anual, multianual a diferentes horizontes, por ciudad, todo el país)

- . Tablas de contingencias precipitación, con los índices aquí listados
- . Scatterplot Temperatura, Error medio (bias), Error cuadrático medio
- . Probabilidad de detección de Tmin, Tmax, (dicotómicas) con scores (cuando la T pronosticada está dentro de un intervalo de $\pm 2^{\circ}\text{C}$).
- . Mapas de índices.
- . Intervalos de confianza

6.7 Pronóstico semanal con GEFS (DNPSS/DSS)

Este pronóstico por el momento no tiene presentación pública pero apunta a objetivos similares al pronóstico por análogos presentado en el punto anterior. Por esta razón es recomendable que ambos pronósticos se verifiquen con un grupo de herramientas comunes.

Recomendación preliminar: *Productos de verificación* (estacional, anual, multianual a diferentes horizontes, por ciudad, todo el país)

- . Tablas de contingencias precipitación, con los índices aquí listados
- . Scatterplot Temperatura, Error medio (bias), Error cuadrático medio
- . Probabilidad de detección de Tmin, Tmax, (dicotómicas) con scores (cuando la T pronosticada está dentro de un intervalo de $\pm 2^{\circ}\text{C}$).
- . Mapas de índices.
- . Intervalos de confianza

6.8 Alertas meteorológicas (DNPSS/DPTA/CPR)

Para el tema específico de alertas un par de referencias útiles, --aunque no muy actuales dada la dinámica de este áreas--, son Sharpe (2010) y Stephenson y otros (2010). Actividades anteriores de verificación de Alertas en el SMN se describen en Suaya y Valdivieso (2018).

En el documento de la OMM *WMO Guidelines on Multi-hazard Impact-based Forecast and Warning Services* (WMO 2015, sección 4.5) se presentan muy brevemente recomendaciones para la verificación por impacto, y se puede observar claramente la falta de metodologías establecidas. Ver también NOAA (1998). Como se discutió en la sección 6.3 de los ACPs, en el caso de las alertas también hay dificultades para definir Aciertos negativos (ver Apéndice C), y por esta razón es de interés que ambos pronósticos mantengan elementos en común.

Recomendación preliminar: *Productos de verificación* (estacional, anual, multianual a diferentes horizontes (si los hay), todo el país)

- . Revisión de categorías de eventos a verificar y umbrales
- . Tabla de contingencia de ACP para toda la República Argentina
- . Utilización de scores que no contabilicen la categoría “aciertos negativos”, no recomendable en estos casos: Bias en la frecuencia, Probabilidad de detección, Falsas alarmas, y Threat score (ver sección 4.3 y <https://www.cawcr.gov.au/projects/verification/>).
- . Estratificación por fenómenos reportados: Tormentas fuertes, lluvias intensas, precipitaciones persistentes, vientos fuertes, etc.
- . Estratificación por región, al menos en acumulaciones trimestrales o anuales
- . Profundización de metodologías para relevamiento por impactos
- . Intervalos de confianza

6.9 Pronósticos de olas de calor y frío

Desde hace tres años que se emiten en el SMN pronósticos por olas de calor durante el verano (ver Herrera y otros 2018), y se espera que para el invierno de 2021 se encuentre operativa la alerta por ola de frío siguiendo una metodología similar. La mencionada Nota Técnica describe una verificación realizada durante el primer año de su operacionalización, pero esta no tuvo desarrollos posteriores.

6.10 Resumen del plan preliminar de verificación

<i>Tipo de pronóstico</i>	<i>Insumos de verificación</i>	<i>observaciones</i>	<i>referencia</i>
Para los próximos 7 días (DNPSS/CPR)	Pronósticos humanos públicos,	Estaciones SMN, estaciones alternativas	WRF, GFS, persistencia, climatología
Muy corto plazo (DNPSS/CPI)	Pronósticos humanos públicos	Estaciones SMN, estaciones alternativas, redes sociales, medios	No hay
Numericos (DCIPS/DMSR)	WRF	Estaciones SMN, estaciones alternativas, sondeos, reanálisis	GFS
Aeronáutico (DMA)	TAF	METAR	WRF, GFS
Estacionales (DNCIPS/DCMC)	Pronósticos humanos, Pronósticos automáticos	Estaciones SMN, productos satelitales	climatología

Pronóstico semanal-análogos (DNCIPS/DCMC)	Pronósticos automáticos	Estaciones SMN	Climatología, pronóstico GEFS
Pronóstico semanal GEFS (DNPSS/DSS)	Pronósticos automáticos	Estaciones SMN	Climatología, pronósticos por análogos
Alertas (DPSS/DOA)	Pronósticos humanos públicos,	Estaciones SMN, estaciones alternativas, redes sociales, medios	No hay

APÉNDICE A: Intervalos de confianza

La estimación de índices de verificación no solo tiene como misión establecer la calidad de una serie de pronósticos, sino también comparar la performance de estos con respecto a otros y permitir además visualizar su evolución en el tiempo. Una propiedad fundamental de estas verificaciones es, entonces, que permitan la comparación. Por esta razón las estimaciones de parámetros como los salidos de un proceso de verificación necesitan de medidas de incertidumbre asociadas.

El objetivo de estas medidas de incertidumbre es que nos ayuden a percibir cuando estos índices se modifican por razones sustanciales --atribuibles a factores específicos, como puede ser la mejora de un producto, la pérdida de una red de medición o la jubilación de un pronosticador estrella-- y no por meras oscilaciones atribuibles al azar. Por ejemplo, que los índices de verificación del WRF cambien de mes en mes no necesariamente a causa de cambios de modelo, sino que puede ser simplemente el resultado de la variación interanual de los campos a pronosticar. La tesitura del pensamiento estadístico es: no hay razón objetiva para que estos índices difieran salvo por las contingencias de mes en mes, por lo tanto debemos diseñar herramientas que sean lo más insensibles posible a estas variaciones.

La situación es un poco más compleja cuando hablamos de pronósticos subjetivos ya que a las variaciones azarosas de la variedad interanual se suman las variaciones en el personal que hace el pronóstico (desde la rotación del personal, al estado de salud, al acceso a la información disponible, a la perturbación de algún elemento del sistema de pronóstico). Esto hace que uno espere que los índices del pronóstico oficial del SMN sufran más variabilidad que los de pronóstico objetivos.

La estimación de los intervalos de confianza puede en general hacerse de dos maneras. La primera, que fue utilizada durante muchas décadas hasta la aparición de computadoras lo suficientemente rápidas, es el cálculo analítico. Es decir, una vez que conocemos el índice que deseamos utilizar, a través de un desarrollo analítico más o menos simple obtenemos el intervalo de confianza. La segunda, utiliza la “fuerza bruta” del poder de cálculo computacional para establecer este intervalo de confianza, basado no sólo en el índice de interés sino también haciendo uso de las muestras obtenidas.

Ambos métodos tienen sus talones de Aquiles, pero en general podemos decir que para ciertos índices (como la estimación de un bias), la solución analítica es simple y de alta calidad, mientras que para otros los métodos numéricos son los únicos disponibles.

Existe mucha información dispersa dedicada a los intervalos de confianza. Un texto accesible que se ocupa del tema es Wilks (2006), y una versión más específica es Gilleland (2010). Si bien estos textos son muy útiles para los métodos analíticos, dedican poco espacio a los métodos numéricos como el “bootstrap”. En esos casos lo mejor resulta el acercarse a las librerías de los lenguajes computacionales de interés. Entre ellos podemos mencionar por ejemplo, el caso de Python (<https://machinelearningmastery.com/calculate-bootstrap-confidence-intervals-machine-learning-results-python/>), y el caso de R (<https://www.statmethods.net/advstats/bootstrapping.html>).

A.1 El intervalo de confianza de Wilson modificado

Los índices asociados a las tablas de contingencia tienen muy variada complejidad, algunos siendo fácilmente resueltos con una fórmula analítica mientras que otros necesitan de bootstrapping. Dada la importancia del caso presentamos a continuación una expresión analítica del intervalo de confianza de alta precisión y mucha utilidad.

En el caso de scores representados como una proporción binomial, una excelente aproximación al intervalo de confianza --llamado *Wilson modificado por continuidad*-- podemos escribirla como

$$s_{inf} = \frac{2ns + z^2 - [z\sqrt{4ns(1-s) + z^2 - \frac{1}{n} + (4s-2) + 1}]}{2(n + z^2)}$$

$$s_{sup} = \frac{2ns + z^2 + [z\sqrt{4ns(1-s) + z^2 - \frac{1}{n} - (4s-2) + 1}]}{2(n + z^2)}$$

Donde s es la proporción a la que le estamos generando el intervalo de confianza, n es el número de la muestra para ese cálculo (usualmente el denominador del cociente) y z es el cuantil $1 - \alpha/2$ de la distribución normal. Para el caso del 95% de confianza, $z = 1.96$. Si s_{inf} resulta menor que cero, se debe tomar $s_{inf}=0$, si s_{sup} es mayor que 1, se debe tomar $s_{sup}=1$. Para una discusión más detallada ver Wallis (2013).

Entre los índices cuyos intervalos de confianza pueden ser estimados de esta manera tenemos la Proporción correcta, la Probabilidad de detección, la Proporción de falsas alarmas, y el Threat score.

APÉNDICE B: Algunos índices muy utilizados para clasificar tablas de contingencia

Una lista completa con referencias de índices de verificación puede encontrarse en <https://www.cawcr.gov.au/projects/verification/>. En lo que sigue se presentan los índices recomendados para nuestras actividades en lo que respecta a tablas de contingencia.

Proporción correcta (accuracy):

$$PC = (\text{Aciertos} + \text{Aciertos negativos}) / \text{total pronosticados}$$

Rango del score: 0 a 1, perfecto: 1.

Bias en la frecuencia (bias score):

$$BIAS = (\text{Aciertos} + \text{Falsas alarmas}) / (\text{Aciertos} + \text{Sorpresas})$$

Rango del score: 0 a ∞ , perfecto: 1.

Probabilidad de detección (hit rate):

$$PD = \text{Aciertos} / (\text{Aciertos} + \text{Sorpresas})$$

Rango de score : 0 a 1, perfecto: 1.

Proporción Falsas alarmas (false alarm ratio):

$$FAR = \text{Falsas Alarmas} / (\text{Falsas Alarmas} + \text{Aciertos})$$

Rango de score : 0 a 1, perfecto: 0.

Threat score:

$$TS = \text{Aciertos} / (\text{Falsas Alarmas} + \text{sorpresas} + \text{Aciertos})$$

Rango de score : 0 a 1, perfecto: 1.

Equitable Threat score:

$$ETS = (\text{Aciertos} - \text{Aciertos al azar}) / (\text{Falsas Alarmas} + \text{sorpresas} + \text{Aciertos} - \text{Aciertos al azar}),$$

donde $\text{Aciertos al azar} = (\text{Aciertos} + \text{sorpresas}) \times (\text{Aciertos} + \text{falsas alarmas}) / \text{total}$.

Rango de score : -1/3 a 1, perfecto: 1. El valor 0 indica ninguna habilidad.

APÉNDICE C: Sobre los “correctos negativos”

La categoría “correcto negativo” en una tabla de contingencia define aquellos pronósticos que pronosticando la no ocurrencia de un evento resultan acertados. Hasta aquí esta categoría no se diferencia de las otras cuatro presentadas en la tabla de la sección 4.2.

Sin embargo, esta categoría no está bien definida para muchos tipos de pronósticos, particularmente aquellos relacionados con las alertas. Este es el caso porque simplemente no hay pronósticos de “no alerta”. Evidentemente uno podría decir que cada pronóstico que no contiene una alerta puede ser considerado un pronóstico de “no alerta”, y esto sucede en muchas instituciones que verifican las emisiones de sus alertas. Sin embargo esto no es siempre posible. Por ejemplo, los ACPs pueden ser emitidos en cualquier momento y por una duración de 3 horas; ¿cómo podemos definir la “no emisión” de ACPs?

Este problema es complejo y diferentes centros lo han resuelto de manera distinta. La Tabla C.I muestra la perspectiva de diferentes autores y centros.

Tabla C.I: Diferentes posiciones con respecto a los “aciertos negativos” a la hora de considerarlos en la verificación de la performance del pronóstico.

Reporte	Postura
WMO 2009	“Since the forecast valid period is 24h, one correct negative should be recorded for each predefined forecast region where no extremes are observed and none forecast. Example: Zimbabwe uses 6 predefined forecast regions. If severe weather is observed in one of these regions, and none is forecast or observed in the other 5, then 5 correct negatives are recorded. If severe weather is forecast for one region and it occurs in another, then a false alarm, a missed event and 4 correct negatives are recorded for that day. Correct negatives are problematic in verification of severe weather; therefore it is acceptable to ignore them altogether, especially if there are relatively few occurrences during the verification period.”
WMO 2010b	“Some NMHSs defined each day as an event, so that the total sample size will equal the number of days in the verification period. For these countries, the contingency table will include all 4 boxes, since inactive days are counted as correct negatives. Other NMHS’s considered only the active days as events, so that their contingency table includes only 3 of the boxes (hits, misses and false alarms) and their sample size is equal to the number of active days.”
Stephenson y otros 2010	“Rather than consider mutual timings of warnings and observed events, weather warnings are usually evaluated by counting the number of compound events: Because estimates of d are either unreliable or unavailable, verification measures for warnings are often limited to those measures based only on a, b, c . The current measures used for verification of severe weather warnings at the Met Office are: <i>Hit rate:</i> $H = a / (a + c)$ (a.k.a. probability of detection) <i>False alarm ratio:</i> $FAR = b / (a + b)$ <i>Threat score:</i> $TS = a / (a + b + c)$ <i>Frequency bias:</i> $B = (a + b) / (a + c)$ ”
Wilson y Giles 2013	“This quantity is often hard to estimate because of the difficulty of establishing bounds on the location and time of a non-occurrence of the event. When nothing is happening weather-wise, questions such as ‘How often is nothing happening?’, ‘When does the non-event start and end?’ etc. arise and are difficult to answer objectively. While establishing the number of occurrences of the non-event may be hard, the forecast of the non-event can be assumed to be the absence of a forecast of the event. It is still necessary to decide how often the non-event is forecast over a verification period. ... This means it is necessary to adopt some form of spatial and temporal discretization of the entire verification period over all regions. ...For convection, a characteristic length of 3 h was selected, and it was assumed that there would be a possibility of occurrence of a severe convective event during 7seven of the eight 3-h periods <i>per day</i> .”
Brooks y Correia 2018	“A challenge exists in defining the correct forecasts of nonevents that are considered in the decision model. We choose to follow Brooks (2004), in setting that value such that the fraction of tornadoes to the total number of events being considered in warning is 0.1. This value is arbitrary and affects the results quantitatively, but not qualitatively...”

Tres de los scores sugeridos por Stephenson y otros (2010), a saber: la Probabilidad de detección PD, la Proporción de falsas alarmas FAR, y el Bias, pueden graficarse en un diagrama muy instructivo llamado “Performance diagram” (ver Roebber 2009). Usando las definiciones se puede ver que se puede graficar como muestra la Fig. C.I.

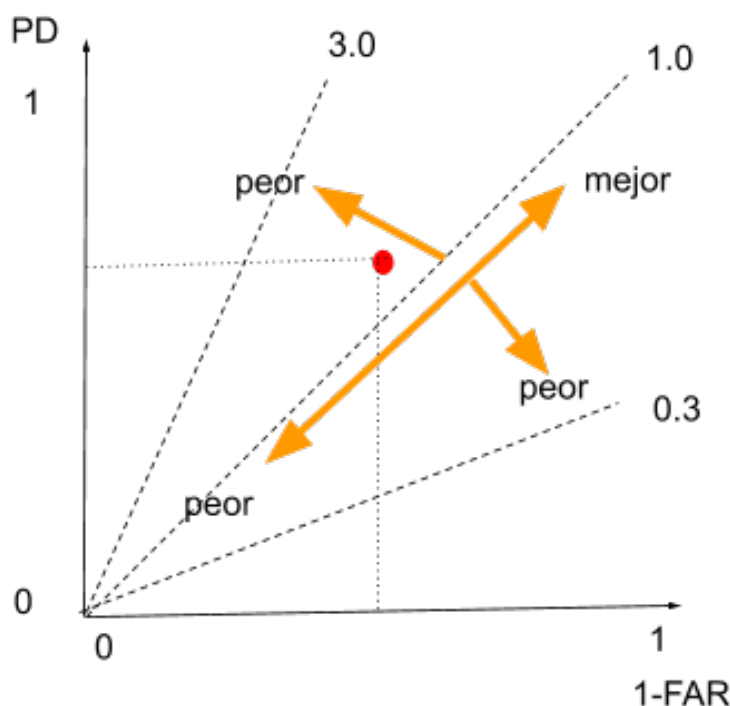


Figura C.I: Ejemplo de diagrama de performance. La abscisa corresponde a 1-FAR, es decir su máximo en 1 implica la ausencia de falsas alarmas. El eje vertical corresponde a la Probabilidad de detección. Las líneas segmentadas corresponden a diferentes valores de Bias, con Bias=1 en la diagonal del gráfico. El resultado ideal corresponde al caso de estudio en el extremo superior derecho, con alta probabilidad de detección, sin falsas alarmas y sin bias. El punto rojo muestra un caso genérico.

En este mismo diagrama se puede graficar el Threat Score (definido en el Apéndice B). Un ejemplo del aspecto final del diagrama puede verse en la Figura C.II. Las curvas correspondientes al Threat Score aparecen como líneas cóncavas. Es interesante notar cómo un mismo score de TS puede obtenerse con muchas o pocas falsas alarmas, probabilidad de detección o bias. Esto es una típica propiedad de los pronósticos categóricos en que la decisión de este balance tiene que ver más con una necesidad del usuario que con las propiedades del sistema de predicción.

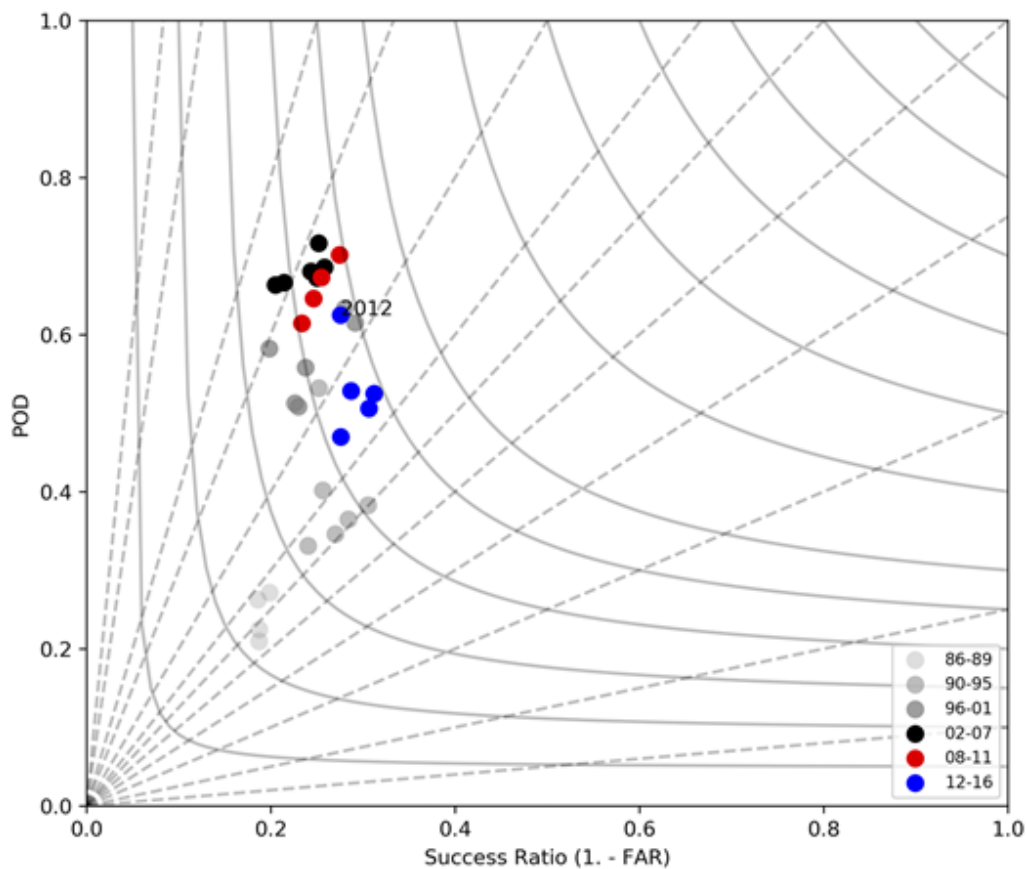


Figura C.II: Ejemplo tomado de Brooks y Correia (2018) del típico diseño del Diagrama de performance. En este caso se pueden ver la performance de alertas por tornado durante 30 años en el NWS. Notar la caída de PODs en la última década estudiada, pero la disminución de falsas alarmas. El equilibrio entre PODs y falsas alarmas es una decisión muy importante que debe dar la institución y depende en gran medida del usuario correspondiente.

7. REFERENCIAS

- Brooks, H. E., and J. Correia, 2018: Long-Term Performance Metrics for National Weather Service Tornado Warnings. *Wea. Forecasting*, **33**, 1501–1511, <https://doi.org/10.1175/WAF-D-18-0120.1>
- Cattani D., A. Faes, M. Giroud Gaillard, M. Matter, 2015: COMFORT: Continuous MeteoSwiss ForecastQuality Score, Scientific Report MeteoSwiss, 99, 45 pp.
<https://www.meteoswiss.admin.ch/content/dam/meteoswiss/en/Ungebundene-Seiten/Publikationen/Scientific-Reports/doc/sr99cattani.pdf>
- Gilleland E., 2010: Confidence Intervals for Forecast Verification. NCAR Technical Note TN-479.
<https://ral.ucar.edu/~ericg/Gilleland2008.pdf>
- Herrera N., Skansi M.M., Beron M.A., Campetella C., Cejas A., Chasco J., Chesini F., de Titto E., Gatto M., Saucedo M., y Suaya M., 2018: Sistema de Alerta Temprana por Olas de Calor y Salud (SAT-OCS). Nota Técnica SMN 2018-50. <http://repositorio.smn.gob.ar/handle/20.500.12160/772>
- Horel J., 2008: Data Representativeness: Are you observing what you want to observe or what you need to observe? https://www.nws.noaa.gov/oh/rfcdev/docs/JHorel_verification_RFCworkshop_Nov08.pdf
- Lopez, P., G. Ryu, B. Sohn, and L. Davies, 2011: Specification of rain gauge representativity error for data assimilation, technical report 647, Eur. Cent. for Medium-Range Weather Forecasts, Shinfield Park, Reading, U. K. <https://www.ecmwf.int/sites/default/files/elibrary/2011/10801-specification-rain-gauge-representativity-error-data-assimilation.pdf>
- Mason S., 2018: Guidance on Verification of Operational Seasonal Climate. WMO 1220. https://library.wmo.int/doc_num.php?explnum_id=4886
- Murphy, A. H. 1973: A new vector partition of the probability score, *J. Appl. Meteorol.*, 12(4), 595–600 [https://doi.org/10.1175/1520-0450\(1973\)012<0595:ANVPOT>2.0.CO;2](https://doi.org/10.1175/1520-0450(1973)012<0595:ANVPOT>2.0.CO;2)
- Murphy, A. H., 1998: The Early History of Probability Forecasts: Some Extensions and Clarifications. *Wea. Forecasting*, 13, 5–15, <https://doi.org/10.1175>
- Mahringer G., 2008: Terminal aerodrome forecast verification in Austro Control using time windows and ranges of forecast conditions. *Meteorological Applications*. 15. 113 - 123.
<https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/met.62>
- NOAA, 1998: NWS's Verification System for Severe and Hazardous Weather Forecasting Needs Modernization. Inspection Report No. IPE-9255.
<https://www.oig.doc.gov/OIGPublications/NOAA-IPE-09255-01-1998.pdf>
- Nurmi, P., 2003: Recommendations on the verification of local weather forecasts. ECMWF Tech. Memo. 430, 18 pp.
<https://www.ecmwf.int/en/elibrary/11401-recommendations-verification-local-weather-forecasts>

OACI, 2007: Servicio meteorológico para la navegación aérea internacional. Anexo 3 al Convenio sobre Aviación Civil Internacional 16ava edición. https://www.wmo.int/pages/prog/www/ISS/Meetings/CT-MTDCF-ET-DRC_Geneva2008/Annex3_16ed.pdf

Ou M.H., M. Charles, and D.C. Collins, 2016: Sensitivity of Calibrated Week-2 Probabilistic Forecast Skill to Reforecast Sampling of the NCEP Global Ensemble Forecast System. *Wea. Forecasting*, **31**, 1093–1107, <https://doi.org/10.1175/WAF-D-15-0166.1>

Pielke R. A., 2001: Further Comments on “The Differentiation between Grid Spacing and Resolution and Their Application to Numerical Modeling”. *Bull. Amer. Meteor. Soc.*, **82**, 699–700, [https://doi.org/10.1175/1520-0477\(2001\)082<0699:FCOTDB>2.3.CO;2](https://doi.org/10.1175/1520-0477(2001)082<0699:FCOTDB>2.3.CO;2).

Roebber, P. J., 2009: Visualizing Multiple Measures of Forecast Quality. *Wea. Forecasting*, **24**, 601–608, <https://doi.org/10.1175/2008WAF2222159.1>.

San Martino F., S. Pérez y G. Russián, 2019: Verificaciones de pronósticos a muy corto plazo en el SMN. Nota Técnica SMN 2019-64. <http://repositorio.smn.gov.ar/handle/20.500.12160/1241>

Sharpe M., 2010: Verification of Weather Warnings. MetOffice. CBS Coordination Group on Forecast Verification (CG-FV), Montreal, Canada, January 2011. https://www.wmo.int/pages/prog/www/DPFS/Meetings/CG-FV_Montreal2011/Doc6-1-Annex.pdf

Stephenson D.B, I.T. Jolliffe, C.A.T. Ferro, C.A. Wilson, M. Sharpe, M. Mittermaier, and T.D. Hewson, 2010: White Paper Review on the Verification of Warnings, Met Office Forecasting Research Technical Report No. 546, 22 pp. <https://library.metoffice.gov.uk/Portal/DownloadImageFile.ashx?objectId=465>

Suaya M y R. Valdivieso, 2018: Verificación de pronósticos de lluvias y alertas meteorológicas en la ciudad de buenos aires: estrategias para el cambio Congremet 2018, Rosario. http://cenamet.org.ar/congremet/wp-content/uploads/2018/09/T267_Suaya.pdf

Taylor K., 2005: Taylor Diagram Primer. https://pcmdi.llnl.gov/staff/taylor/CV/Taylor_diagram_primer.pdf?id=96

Wallis S., 2013: Binomial Confidence Intervals and Contingency Tests: Mathematical Fundamentals and the Evaluation of Alternative Methods, *Journal of Quantitative Linguistics*, 20:3, 178-208, DOI: 10.1080/09296174.2013.799918

Wilks D.S., 2006: Statistical Methods in the Atmospheric Sciences. 2nd Edition, Academic Press, London.

Wilson L., A. Giles, 2013 : A new index for the verification of accuracy and timeliness of weather warnings. *Meteorol. Appl.* 20, 206–216. <https://rmets.onlinelibrary.wiley.com/doi/epdf/10.1002/met.1404>

WMO, 2000: Guidelines on Performance Assessment of Public Weather Services, TD No. 1023. http://www.wmo.int/pages/prog/hwarp/documents/FFI/expert/Guidelines_on_Performance_Assessment_of_Public_Weather_Services.pdf

WMO, 2004: Recommendations for the Verification and Intercomparison of QPFs and PQPFs from Operational NWP Models. https://cawcr.gov.au/projects/verification/WGNE/QPF_verif_recomm.pdf

WMO, 2009: SWFDP - Southern Africa QUARTERLY PROGRESS REPORT N° 2 for the period 1 April 2009 – 30 June 2009.

http://www.wmo.int/pages/prog/www/DPFS/Meetings/SG-SWFDP_Geneva2010/documents/doc3-2.doc.

WMO, 2010a: Manual on the Global Data-processing and Forecasting System, Annex IV to the WMO Technical Regulations. WMO-No. 485.

https://www.wmo.int/pages/prog/www/DPFS/documents/485_Vol_I_en.pdf

WMO, 2010b: Steering group severe weather forecasting demonstration project (SWFDP), Geneva, Switzerland, 23 - 26 February 2010. https://www.wmo.int/pages/prog/www/DPFS/Meetings/ET-OWFPS_Beijing2018/documents/DOC-3-2_review_actions_last_meeting.doc

WMO, 2015: *WMO guidelines on multi-hazard impact-based forecast and warning services*. WMO TD no. 1150. https://library.wmo.int/doc_num.php?explnum_id=7901

WMO, 2017: Guidelines for Nowcasting Techniques. WMO-No. 1198.

https://library.wmo.int/doc_num.php?explnum_id=3795

Zingerle, C. and Nurmi, P., 2008, *Monitoring and verifying cloud forecasts originating from operational numerical models*, Meteorological Applications, 15(3), 325–330. https://www.researchgate.net/figure/Double-penalty-effect-A-high-resolution-forecast-of-the-same-pattern-as-the-observations_fig6_227934533

Instrucciones para publicar Notas Técnicas

En el SMN existieron y existen una importante cantidad de publicaciones periódicas dedicadas a informar a usuarios distintos aspectos de las actividades del servicio, en general asociados con observaciones o pronósticos meteorológicos.

Existe no obstante abundante material escrito de carácter técnico que no tiene un vehículo de comunicación adecuado ya que no se acomoda a las publicaciones arriba mencionadas ni es apropiado para revistas científicas. Este material, sin embargo, es fundamental para plasmar las actividades y desarrollos de la institución y que esta dé cuenta de su producción técnica. Es importante que las actividades de la institución puedan ser comprendidas con solo acercarse a sus diferentes publicaciones y la longitud de los documentos no debe ser un limitante.

Los interesados en transformar sus trabajos en Notas Técnicas pueden comunicarse con Ramón de Elía (rdelia@smn.gov.ar), Luciano Vidal (lvidal@smn.gov.ar) o Martin Rugna (mrugna@smn.gov.ar) de la Dirección Nacional de Ciencia e Innovación en Productos y Servicios, para obtener la plantilla WORD que sirve de modelo para la escritura de la Nota Técnica. Una vez armado el documento deben enviarlo en formato PDF a los correos antes mencionados. Antes del envío final los autores deben informarse del número de serie que le corresponde a su trabajo e incluirlo en la portada.

La versión digital de la Nota Técnica quedará publicada en el Repositorio Digital del Servicio Meteorológico Nacional. Cualquier consulta o duda al respecto, comunicarse con Melisa Acevedo (macevedo@smn.gov.ar).