MODELO ESTADÍSTICO PARA LA PREDICCIÓN DEL ÍNDICE ESTANDARIZADO DE SEQUÍA PLUVIOMÉTRICA (IESP) EN ANDALUCÍA

Rafael BLANQUERO[1], Emilio CARRIZOSA[1], Mª Fernanda PITA[2],
Juan Mariano CAMARILLO[2], José Ignacio ÁLVAREZ-FRANCOSO[2]
[1]*Depto. de Estadística e Investigación Operativa. Universidad de Sevilla*
[2]*Depto. de Geografía Física y AGR. Universidad de Sevilla*
rblanquero@us.es, ecarrizosa@us,.es, mfpita@us.es,
jmcamarillo@us.es, jalvarez2@us.es

RESUMEN

La comunicación aborda el diseño de un modelo estadístico de predicción dinámica de la sequía en Andalucía y su persistencia en un horizonte temporal de 12 meses a partir de los datos históricos (1950-2012) del Índice Estandarizado de Sequía Pluviométrica (IESP) en 243 observatorios de Andalucía. Se emplea un algoritmo kNN (k-*Nearest Neighbors*), que busca las situaciones pasadas más similares a la actual y predice el futuro promediando lo que ocurrió a continuación en dichas situaciones. Los resultados producen porcentajes de error muy reducidos. El modelo se está aplicando en rutina en el Sistema de Información de Climatología Ambiental (CLIMA) de la Consejería de Medio Ambiente de la Junta de Andalucía (http://www.climasig.es).

Palabras clave: indicador de sequía, predicción de sequía, IESP, k-Nearest Neighbors, Andalucía

ABSTRACT

This paper presents a statistical model for the dynamical prediction of drought in Andalusia. The drought index used is the IESP, which stands for Índice Estandarizado de Sequía Pluviométrica in Spanish, and has been applied to historical observation series (1950-2012) for 243 observatories. A kNN (k-Nearest Neighbors) algorithm is used to predict the future situations by averaging the past instances that are most similar to the current one. The application of a validation procedure on the available data shows that this model leads to very small percentages of failed predictions. The model is being applied by the Environment Ministry of the Andalusian Government to monitor drought in the region.

Key words: drought indicator, drought prediction, IESP, k-Nearest Neighbors, Andalusia

1. INTRODUCCIÓN

En el análisis de la sequía la tarea prioritaria es la de encontrar un indicador que refleje adecuadamente el estado de sequía existente en cada momento en una región, pero, de cara a la gestión de los recursos hídricos, junto a este diagnóstico de la situación se requiere igualmente una

predicción del estado que adoptará la sequía en el futuro inmediato, dado que esta situación futura determinará en buena medida el abanico de medidas a adoptar.

Mientras que el análisis y diagnóstico han sido muy abordados y hay numerosas propuestas de indicadores de sequía adaptados a diferentes aplicaciones, no ocurre lo mismo en relación con la predicción del futuro para la sequía, aspecto para el que no se encuentran referencias tan abundantes, siendo todavía más numerosos los intentos de predicción de las precipitaciones futuras. Éstas se llevan a cabo, bien a partir de los propios datos pasados de las precipitaciones (Azadi y Sepaskhah, 2011), bien a partir de éstos más otros indicadores climáticos (ENSO, NAO…) en aquellos ámbitos en los que se ha puesto en evidencia de manera clara la relación existente entre ambos (Kulkarni *et al*, 2012; Nair *et al*, 2012; Rivera *et al*, 2012).

Hablando de la predicción de la sequía en términos estrictos, existen ya interesantes revisiones que muestran la diversidad de aproximaciones existentes en relación con el tema (Maier y Dandy, 2000; Goddard *et al*, 2001; Mishra y Singh, 2011). Los métodos estadísticos tradicionales arrojan muy buenos resultados a partir de la consideración, bien de los índices de sequía antecedentes y/o de las precipitaciones subyacentes a la elaboración de esos índices (Mishra y Desai, 2005; Cancelliere *et al*, 2007; Araghinejad, 2011), bien de la consideración de estos inputs más la adición de predicciones climáticas (Hwang y Carbone, 2009). Junto a ellos cada vez se utilizan más para la predicción de la sequía y de otras variables hidrológicas las redes neuronales en sus diferentes aproximaciones (Dawson y Wilby, 2001; Mishra y Desai, 2006; Morid *et al*, 2007).

Nuestro trabajo se centra precisamente en este aspecto de la sequía, la predicción del futuro. De hecho, el objetivo de la comunicación es el diseño de un modelo estadístico de predicción dinámica de la persistencia temporal de la sequía en Andalucía con un horizonte temporal de R=12 meses a partir de los valores mensuales del Índice Estandarizado de Sequía Pluviométrica (IESP). Dicho modelo de predicción estaría destinado a integrarse en el sistema de seguimiento de la sequía, integrado a su vez en el Subsistema de Información de Climatología Ambiental (CLIMA) de la Consejería de Medio Ambiente de la Junta de Andalucía (http://www.climasig.es).

2. EL ÍNDICE ESTANDARIZADO DE SEQUÍA PLUVIOMÉTRICA (IESP)

El IESP es un índice mensual de sequía pluviométrica que se basa en el cálculo de las anomalías pluviométricas mensuales acumuladas, de modo similar al muy conocido Standardized Precipitation Index (SPI) de Mac Kee. Al igual que en el SPI, los valores del índice son anomalías pluviométricas acumuladas estandarizadas y en él los valores negativos corresponden a meses secos, en tanto que los positivos reflejan meses no secos. El índice responde a la expresión:

$$IESP_i=(APA_i-APA_{med})/\sigma_{APA}$$

donde:

$IESP_i$ = Índice Estandarizado de Sequía Pluviométrica del mes i.

APA_i = Anomalía pluviométrica acumulada del mes i.

APA_{med}= Valor medio de las anomalías pluviométricas acumuladas del mes correspondiente.

σ_{APA}= Desviación típica de las anomalías pluviométricas acumuladas del mes correspondiente.

Por su parte, APA_i responde a la expresión: $APA_i=\Sigma AP_i$

Desde $i=$ 1 hasta $AP_i<0$ y $APA_{i-1}\geq 0$, siendo AP_i la anomalía pluviométrica del mes *i.*

La esencia del índice, y su seña de identidad frente a otros similares, es que reinicia los cálculos de las anomalías acumuladas cada vez que se produce un nuevo mes seco ($APi<0$) en el marco de un

periodo excedentario (con $APA_{i-1} \geq 0$); ello permite reflejar las secuencias secas de diferentes longitudes a partir de una única elaboración del índice, frente al SPI, que requiere una aplicación a múltiples escalas temporales para reflejar las diferentes duraciones de la sequía (Pita, 2001; Pita, 2007).

3. DATOS Y METODOLOGÍA

Para el desarrollo del modelo se han utilizado los datos históricos (1950-2012) del IESP en 243 estaciones de observación de Andalucía, realizándose un análisis independiente para cada una de ellas (ver figura 1).

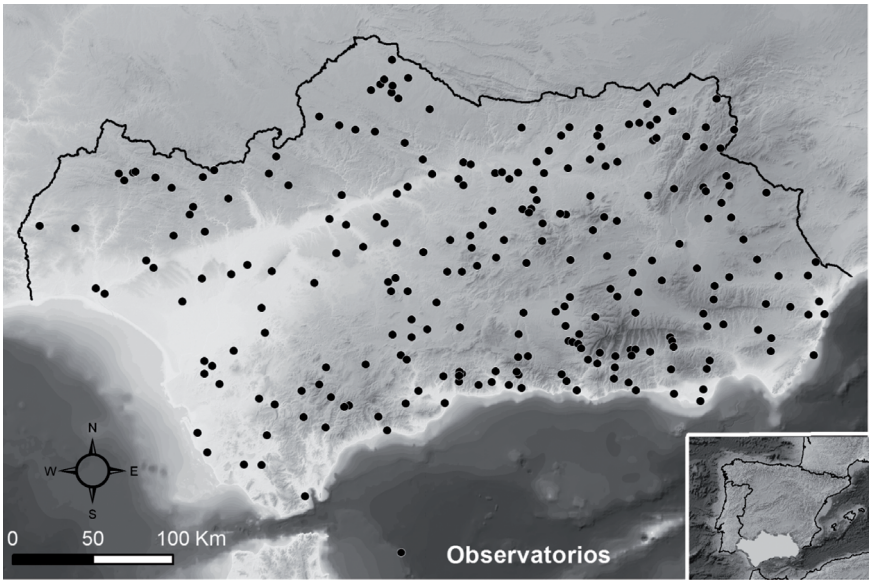

FIG. 1: *Localización de los observatorios estudiados.*

El principal condicionante para la elección del método ha sido la necesidad de que éste sea implementado en tiempo real dentro del sistema CLIMA, de la Red de Información Ambiental de Andalucía (REDIAM). Ello ha impuesto la conveniencia de utilizar un número reducido de variables para la elaboración de las predicciones (solo se usan los valores del IESP de las series), así como un procedimiento de cálculo no excesivamente sofisticado ni exigente en recursos computacionales. Teniendo en cuenta estos condicionantes, se realizó una primera prueba, mediante el software de minería de datos Weka (http://www.cs.waikato.ac.nz/ml/weka/), consistente en predecir el estado de "sequía" o "no sequía" del siguiente periodo (t+1). Los clasificadores utilizados fueron: *Naive Bayes, Regresión logística, Redes neuronales, Árboles de clasificación (algoritmo C4.5), Support Vector Machine y k-Nearest Neighbors,* obteniendo en todos los casos resultados similares.

Razones de sencillez y de facilidad de comprensión e implementación en rutina nos impulsaron a explorar el detalle y, en su caso, a desarrollar este último.

En consecuencia, el método de predicción utilizado se basa en el empleo de un algoritmo kNN (del inglés **k-*N*earest *N*eighbors**) (Fix y Hodges, 1951). Resulta una metodología sobradamente

contrastada en la literatura estadística (Bennayan y Hoogenboom, 2008; Brabec y Meister, 2001; Ouarda *et al*., 1999) y muy adecuada a nuestro caso tanto por su simplicidad como por su capacidad para predecir correctamente fenómenos complejos.

Se puede afirmar que el algoritmo kNN busca las k situaciones pasadas más similares a la actual, y predice el futuro promediando lo que ocurrió a continuación en dichas situaciones. Se trata de un método de clasificación supervisada no paramétrico especialmente empleado en fenómenos de los que *a priori* no se conoce su modelo de distribución, o su ajuste presenta dificultades, como es nuestro caso. Se trabaja, por tanto, con la muestra empírica y se generan predicciones basadas en la similitud entre las cadenas de vecinos más próximos presentes en dicha muestra respecto a la cadena actual objeto de predicción.

El proceso tiene dos fases: la fase de aprendizaje y la de predicción. En la primera se determinan los tres parámetros que se usarán en el modelo: el ancho de ventana (v) o ancho de la cadena óptimo (número de meses que la componen) para establecer las comparaciones entre los k vecinos más similares, el número de vecinos (k) o número de cadenas óptimo para establecer las similitudes con la cadena objeto de predicción, y el peso del mes (w) o valor de la ponderación del último mes que compone la cadena objeto de predicción. Dichos parámetros deben recalcularse una vez al año para permitir la aplicación del método en tiempo real. Una vez validados, dichos parámetros se usarán en la fase de predicción para estimar mensualmente la probabilidad de mantenimiento de la sequía. A continuación se explicitan ambas con más detalle.

3.1. La fase de predicción

Una vez establecidos los tres parámetros (v, k, w) en la fase de aprendizaje, que describiremos más adelante, el algoritmo propuesto busca aquellas k situaciones pasadas más parecidas a la actual y predice el futuro promediando lo que ocurrió a continuación en dichas situaciones pasadas similares. Las predicciones se realizan sobre los siguientes parámetros:

– P1(S): Probabilidad de que el próximo periodo sea de sequía (indicador IESP negativo).

– P2(S): Probabilidad de que los dos próximos periodos sean de sequía (indicador IESP negativo en los dos próximos periodos).

– …

– Pr(S): Probabilidad de que los r próximos periodos sean de sequía (indicador IESP negativo en los r próximos periodos).

El procedimiento de estimación consta de los siguientes pasos:

3.1.a. BASE DE DATOS

Creación de una base de datos en la que los registros son de la forma:

$$t,\ I(t),\ I(t+1),\ I(t+2),\ldots,\ I(t+v-1),$$

donde:

– t es el mes del año.

– $I(t)$ es el valor del IESP en el mes t.

– $I(t+1)$ es el valor del IESP en el mes $t+1$.

– …

– $I(t+v-1)$ es el valor del IESP en el mes $t+v-1$.

En otras palabras, los registros de esta base de datos, que llamaremos *cadenas*, son vectores con v+1 componentes, siendo la primera componente un mes y las siguientes los valores del IESP en los meses t, t+1,…, t+v-1.

3.1.b. Búsqueda de cadenas próximas

Dada la cadena c, buscar en la base de datos el conjunto C de k cadenas más similares o próximas a c, donde la proximidad se mide a partir de la distancia euclídea ponderada. La proximidad *d(t,t')* entre las cadenas *(t, I(t), I(t+1),..., I (t+v–1)) y (t', I(t'), I(t'+1),..., I(t+v–1))* viene dada por:

$$d(t,t')=w\left(s(t)-s(t')\right)^2+\left(I(t)-I(t')\right)^2+\ldots+\left(I(t+v-1)-I(t'+v-1)\right)^2,$$

donde: *s(t)* y *s(t')* representan los valores de *t* y *t'* estandarizados, es decir, transformados primero a la escala *1,2,..., 12* y luego transformados mediante estandarización de forma que la media sea 0 y la desviación típica sea 1.

3.2. La fase de aprendizaje

El objetivo fundamental de esta fase es la estimación de los parámetros (*v,k,w*) para cada estación, de modo que se optimice una medida del error global del proceso de estimación en dicha estación. Para ello se divide la base de datos en dos muestras diferentes: la *muestra de aprendizaje* y la *muestra de validación*. Como suele ser habitual en estos casos, y dado el carácter continuo y acumulativo del índice, que impide seleccionar al azar ambas muestras, la muestra de aprendizaje contiene el 66% de los datos más antiguos y la segunda el 33% de los datos más recientes. Para cada elección de parámetros (*v,k,w*) se evalúa sobre la muestra de validación la regla obtenida si consideramos la muestra de aprendizaje como base para construir las secuencias vecinas, y se calculan las tasas de acierto.

– *n1(S)*: número de secuencias en las que se predice correctamente que el próximo periodo es de sequía.

– *n'1(S)*: número de secuencias en las que se predice correctamente que no ocurre que el próximo periodo es de sequía.

– *n2(S)*: número de secuencias en las que se predice correctamente que los dos próximos periodos son de sequía.

– *n'2(S)*: número de secuencias en las que se predice correctamente que no ocurre que los dos próximos periodos son de sequía.

– …

– *nr(S)*: número de secuencias en las que se predice correctamente que los *r* próximos periodos son de sequía.

– *n'r(S)*: número de secuencias en las que se predice correctamente que no ocurre que los *r* próximos periodos son de sequía.

Las magnitudes *n'* son objeto de una corrección durante el proceso de determinación de las mismas. Dicha corrección está encaminada a conseguir una elección de parámetros que potencie la detección de la finalización de un periodo de sequía, por considerarse éste un hecho de especial relevancia. Así, cuando en una secuencia dada se predice correctamente que no ocurre que los *t* próximos periodos sean de sequía, partiendo de una situación en la que los *t− 1* periodos lo son, *n't(S)* se incrementa en una cantidad adicional *p = 1*.

Finalmente se escogen los valores de *v, k* y *w* para los que se hace máximo el indicador de fiabilidad del método *f(v, k,w)*:

$$f(v,k,w)=n1(S)+n'1(S)+n2(S)+bn'2(S)+n3(S)+b^2n'3(S)+\ldots+nr(S)+b^{r-1}n'r(S)$$

El indicador de fiabilidad, lógicamente, adopta un valor tanto más grande cuanto mayores sean los aciertos; por otro lado, se adopta un coeficiente multiplicativo *b* para los *n'* de 0.9, de forma que se otorgue más peso a los aciertos de sequía que a los de no sequía; además, se ponderan más los aciertos cercanos que los lejanos, de ahí los exponentes progresivos del coeficiente *b*.

Para encontrar los valores óptimos *(v,k,w)* se ha realizado una *búsqueda en rejilla* (véase Pintér, 1996), evaluando f(v,k,w) para una rejilla de valores en los que v y k han variado en el conjunto *{1,2,,...,10}* y w en el conjunto *{1/10, 2/10,...,10/10}*, seleccionándose la terna que aportara mayor valor para *f*.

4. RESULTADOS

4.1. Los parámetros óptimos

Los parámetros óptimos encontrados varían mucho entre los distintos observatorios (ver figura 2). La anchura óptima de la ventana (*v*) se sitúa claramente en los valores bajos. En más del 88% de los observatorios las cadenas iguales o inferiores a 3 meses son las que arrojan mejores predicciones y en casi el 50% de los mismos las mejores predicciones se logran con cadenas de un solo mes. En los vecinos, o número de cadenas necesario para hacer las estimaciones (*k*), la respuesta es más uniforme; todos los tamaños de vecinos registran frecuencias apreciables, destacando las de 1, 3 y 10. En cuanto a la ponderación del último mes de la cadena (*w*), al igual que en el ancho de la ventana, registra un predominio de los valores bajos, destacando especialmente el valor de 0,1.

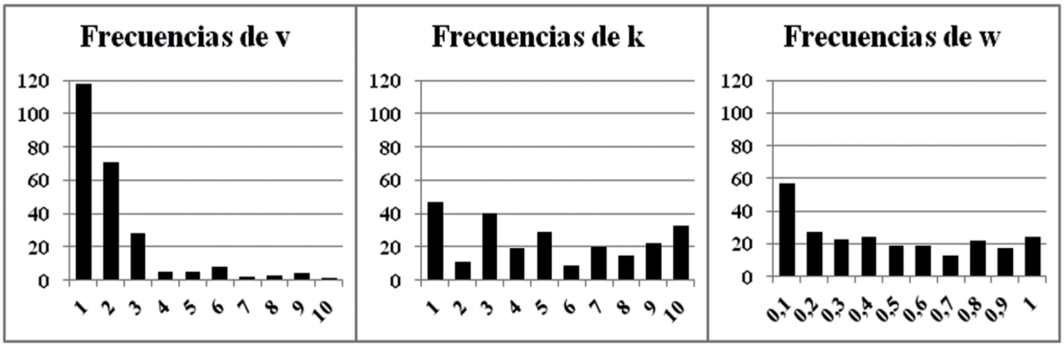

FIG. 2: *Frecuencias registradas por los parámetros óptimos del modelo en los observatorios estudiados. (v = ancho de la ventana, k = número de vecinos, w = factor ponderador del último mes, n=243.*

4.2. La fiabilidad de la estimación

La aplicación del método a los 243 observatorios arroja muy buenos resultados, con un porcentaje medio de fallos de solo 7,91%. El valor mínimo se sitúa en 0% y el máximo asciende a 32,21. Hay que advertir, no obstante, que este valor es absolutamente excepcional; en realidad, la mayoría de los observatorios registran porcentajes de fallos muy bajos: casi el 50% de los mismos registra errores inferiores al 8% y en el 90% de los casos los porcentajes de error son inferiores al 12% (ver figura 3)

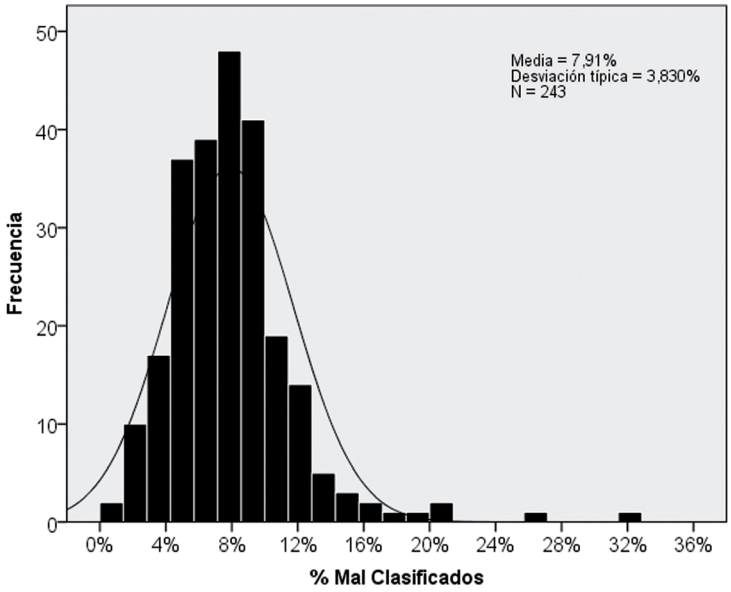

FIG. 3: *Histograma de frecuencias de los porcentajes de error en los observatorios analizados.*

4.3. Variaciones de la fiabilidad en función del alcance temporal de las predicciones.

Los porcentajes de error muestran ciertas heterogeneidades en función del alcance temporal de las predicciones. Los porcentajes medios adoptan siempre valores reducidos, pero éstos son crecientes a medida que se incrementa el alcance temporal de la predicción, superándose el valor de 10 para predicciones con alcance igual o superior a 10 meses (ver figura 4).

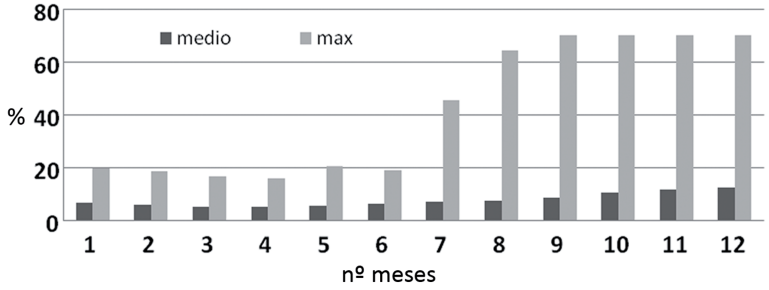

FIG. 4: *Valores medios y máximos de los porcentajes de error en función del alcance temporal de la predicción.*

Estas variaciones son aún más acusadas en los porcentajes de error máximos, los cuales para periodos reducidos nunca alcanzan el 20%, en tanto que para periodos superiores a 9 meses llegan a alcanzar valores que rebasan el 70%. Hay que señalar, no obstante, que estos parámetros se disparan sólo gracias a la intervención de algunos valores excepcionalmente altos que se producen en los casos de alcance temporal elevado, prolongando mucho las colas de los respectivos histogramas de frecuencias; si se prescinde de estos extremos, los porcentajes de error se mantienen en unos valores aceptables incluso en las predicciones de largo alcance (ver figura 5).

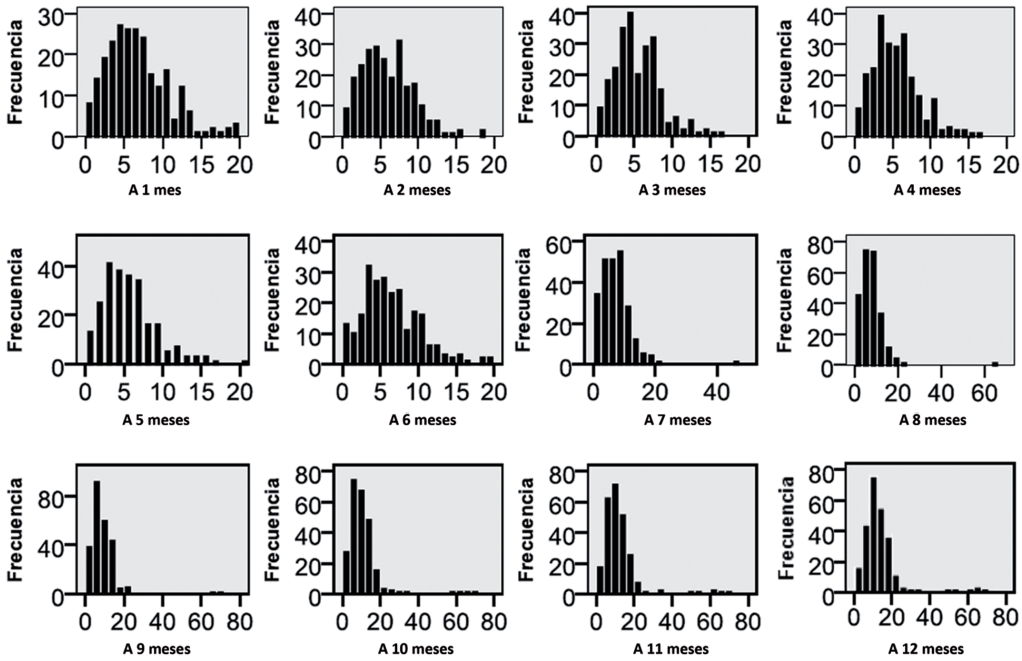

FIG. 5: *Histogramas de frecuencias de los porcentajes de error en función del alcance temporal de las predicciones*

5. DISCUSIÓN Y CONCLUSIONES

El modelo de predicción utilizado es lo suficientemente sencillo como para poder ser desarrollado en rutina y en tiempo real en amplios espacios y para numerosas estaciones de observación, lo que lo convierte en un instrumento útil para su implementación en sistemas de seguimiento y alerta de sequía. En nuestro caso el modelo ha arrojado resultados suficientemente satisfactorios como para permitir su aplicación en tiempo real dentro del Subsistema de Información de Climatología Ambiental (CLIMA) de la Consejería de Medio Ambiente de la Junta de Andalucía. Ello no implica que no haya que realizar mejoras en el modelo, especialmente en las predicciones de mayor horizonte temporal, que son las que arrojan valores máximos más extremos. También se consideran líneas prioritarias de futuro: la predicción de la severidad de la sequía y no solo la probabilidad de su existencia y la posibilidad de inclusión de nuevos parámetros en el modelo, siempre que éstos no dificulten su aplicación en tiempo real. Así mismo se hace necesario un examen detallado de la pauta espacial seguida por la fiabilidad de la predicción, especialmente en los horizontes temporales más largos y en sus variaciones mensuales, que podrían en buena medida estar determinadas por diferencias en los regímenes pluviométricos registrados en las distintas áreas de Andalucía.

Agradecimientos

El trabajo ha sido realizado en el marco de los proyectos: "*Directiva Marco del Agua y Riesgos Hídricos: Gestión y Mitigación de Sequías*", CSO2011-29425 del Plan Nacional de I+D+i, "*Desarrollo de un modelo de anticipación a las sequías basado en escenarios dinámicos (GUADALSEQ)"*, Proyecto de Excelencia de la Junta de Andalucía (HUM-7922) y "*Sustainable Water Action (SWAN). Building research links between EU and USA*" del VII programa Marco de la Unión Europea. Se ha beneficiado de los fondos aportados por la

Consejería de Medio Ambiente de la Junta de Andalucía. Los datos proceden del Subsistema CLIMA, que integra las redes de observación de la AEMET, la Consejería de Medio Ambiente y la Consejería de Agricultura de la Junta de Andalucía.

REFERENCIAS

Araghinejad, S. (2011): "An Approach for Probabilistic Hydrological Drought Forecasting", *Water Resources Management*, 25,191–200

Azadi, S. y Sepaskhah, A.R. (2011): "Annual precipitation forecast for west, southwest, and south provinces of Iran using artificial neural networks", *Theoretical and Applied Climatology*, Published on line, DOI 10.1007/s00704-011-0575-9

Banayan, M. y Hoogenboom, G. (2008): "Weather analogue: a tool for real-time prediction of daily weather data realizations based on a modified k-nearest neighbor approach", *Environmental Modelling and Software*, 23, 703-713

Brabec, B. y Mesister, R. (2001): "A nearest-neighbor model for regional avalanche forecasting", *Annals of Glaciology,* 32, 130-134

Cancelliere, A., Di Mauro, G., Bonaccorso, B. yG. Rossi, G. (2007): "Drought forecasting using the Standardized Precipitation Index", *Water Resources Management*, 21, 801–819

Dawson, C.W. y Wilby, R.L. (2001): "Hydrological modeling using artificial neural networks" , *Progress in Physical Geography*, 25, 80-108

Fix, E., Hodges, J.L., 1951. Discriminatory analysis—nonparametric discrimination: consistency properties. Technical Report 4, USAF School of Aviation Medicine, Randolph Field, TX. Published in: Agrawala, A. (Ed.), Machine Recognition of Patterns. IEEE Press, New York, 1977

Goddard, L., Mason, S.J., Zebiak, S.E., Ropelewski, C.F., Basher, R. y Cane, M.A. (2001): "Current Approaches to Seasonal-to- Interannual Climate Predictions", *International Journal of Climatology,* 21, 1111–1152

Hwang, Y y Carbone, G.J. (2009): "Ensemble Forecasts of Drought Indices Using a Conditional Residual Resampling Technique", *Journal of Applied Meteorology and Climatology,* 48, 1289-1301

Kulkarni, M.A., Acharya, N., Kar, S.C., Mohanty, U.C., Tippett, M.K., Robertson, A.W., Luo, J.J. y Yamagata, T. (2012): " Probabilistic prediction of Indian summer monsoon rainfall using global climate models", *Theoretical and Applied Climatology,* 107, 441–450

Maier, H.R. y Dandy, G.C. (2000): "Neural networks for the prediction and forecasting of water resources variables: a review of modelling issues and applications", *Environmental Modelling& Software* 15, 101–124

Mishra, A.K. y Desai, V.R., (2005): "Drought forecasting using stochastic models". *Stochastic Environmental Research and Risk Assessment*. 19, 326–339.

Mishra, A.K. y Desai, V.R. (2006): "Drought forecasting using feed-forward recursive neural network"**,** *Ecological Modeling,* 198, 127–138

Mishra, A.K. y Singh, V.P. (2011): "Drought modeling – A review", *Journal of Hydrology*, 403, 157–175

Morid, S., Smakhtinb, V. y Bagherzadeh, K. (2007): "Drought forecasting using artificial neural networks and time series of drought indices", *International Journal of Climatology, 27*, 2103–2111

Nair, A., Mohanty, U.C. y Acharya, N. (2012): "Monthly prediction of rainfall over India and its homogeneous zones during monsoon season: a supervised principal component regression approach on general circulation model products", *Theoretical and Applied Climatology*, Published on line, DOI 10.1007/s00704-012-0660-8

Ouarda, T., Hache, M. Rasmussen, P.F. y Bobee, B. (1999): "Generation of ESP forecast series with the non parametric nearest neighbor approach", *Water,* 99, 1154-1159

Pintér, J.D. (1996): "Continuous Global Optimization Software: A Brief Review", *Optima,* 52, 1-8.

Pita, M.F.(2001): "Un nouvel indice de sécheresse pour les domaines méditerranéens. Application au bassin du Guadalquivir (sudouest de l'Espagne)", *Publications de l'Association Internationale de Climatologie*, vol. 13, Nice, pp. 225-233

Pita, M.F. (2007): "Recomendaciones para el establecimiento de un sistema de indicadores para la previsión, el seguimiento y la gestión de la sequía", en Cabrera, E. y Babiano, L.: *La sequía en España. Directrices para minimizar su impacto,* Madrid, Ministerio de Medio Ambiente, pp. 107-132 . Disponible en:
http://hispagua.cedex.es/sites/default/files/hispagua_documento/sequia_espana.pdf

Rivera, D., Lillo, M., Uvo, C.B., Billib, M. y Arumí, J.L. (2012): "Forecasting monthly precipitation in Central Chile: a self-organizing map approach using filtered sea surface temperature", *Theoretical and Applied Climatology*, 107, 1–13