

## WEEK 10

### Deep learning and satellite based precipitation for spatial extreme drought analysis

#### 1. Objective

The objective of next study is improving spatial-tempo extreme drought tracking and prediction. Instead of drought severity, drought duration, drought hot spots were considered, the drought areas, drought propagation, direction will be analyzed. The spatial-temporal drought results support the manage and reduce drought damages. Besides that, we would like to figure out benefits of combine GBP and SBP in spatial drought analysis.

First, merge satellite-based data (SBD) to gauge-based data (GBD) and machine learning (ML) can improve the accuracy of the spatial coverage. We are going to use the method of the study (Meyer et al., 2019) to combine SBD and GBD. Fig1 shows the general idea improve quality of monitors air temperature in South America by using SBD and GBD. Results shows that deep learning, SBD and GBD can be promising method for hydrological spatial analysis.

#### Example: Monitoring air temperature in Antarctica

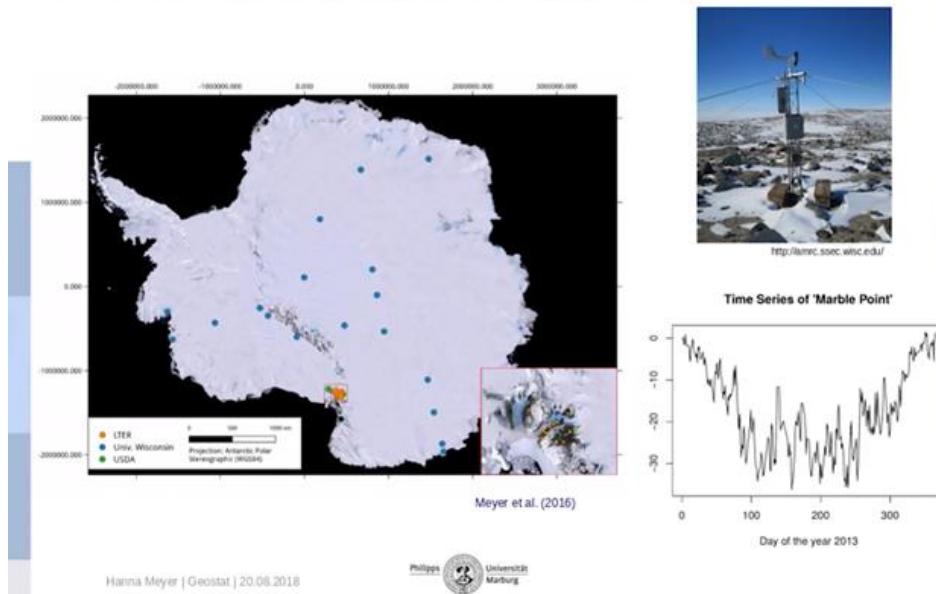


Fig 1. General idea to use SBD and ML for spatial analysis the hydrology (Meyer et al., 2019)

Second, the tempo-spatial analysis are important works to help water resource managers tracking hydrological events. In drought studies, comprehensive extreme drought events should consider drought severity, drought duration, and drought area coverages. In recent, overview about spatiotemporal analysis of extreme drought events was presented in study (Diaz et al., 2019). This study demonstrates spatial analysis extreme drought for Mexico and India as case studies. The spatial-temporal analysis of drought toolbox (STAND) was used to investigate extreme drought events and tracking. Fig 2 shows the concept methodology for estimation the spatial extend drought areas in India.

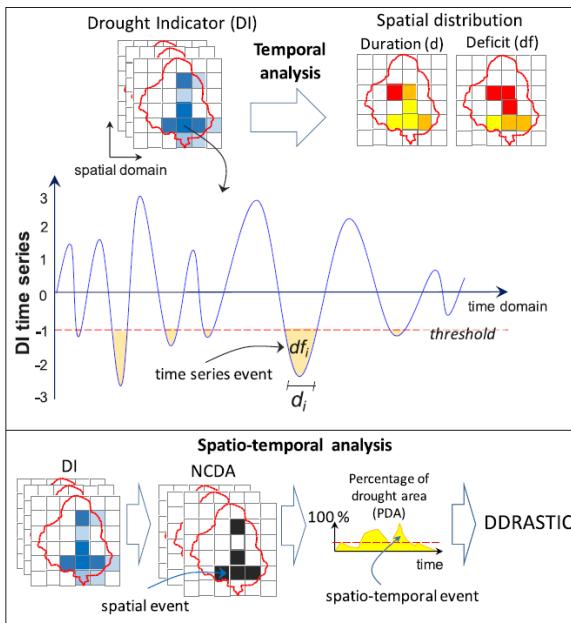


Fig 2. Schematic overview of the methodologies for drought analysis in the STAND toolbox: non-contiguous drought area (NCDA) analysis and drought duration, severity, and intensity computing (DDRASTIC) (Diaz et al., 2019).

## 2. Methodology

### Overview

We literature review the spatial-tempo analysis for extreme drought to setup the methodology. Spatial-temporal drought analysis has been studied using various methods. Drought -severity-area-frequency curve using estimated grid SPI estimated severity and locations for Kangsabati River basin, India (Mishra & Desai, 2005). Clustering algorithms have also been applied many works. For instance, principal component analysis and K-mean clustering to the SPI time series for spatial-temporal pattern of drought in Portugal (Santos et al., 2010). Detail of extreme drought spatial

characteristics was presented in the useful study (Ren et al., 2018). The regional extreme events included drought events was clarified various approaches. It provides the important idea for definition extreme drought events considered spatial characteristic.

#### |Describe the methodology

In this study we propose utilizing spatial track method (Diaz et al., 2020) to determine the temporal-spatial extreme drought events. General spatial tract method (S-track) consist of three main steps: (1) calculation of the spatial drought units (referred to here also as areas or clusters); (2) localization of centroids; and (3) linkage of centroids. Overview S-track method is presented in Fig 3.

**Commented [VQT1]: phương pháp thứ Strack**

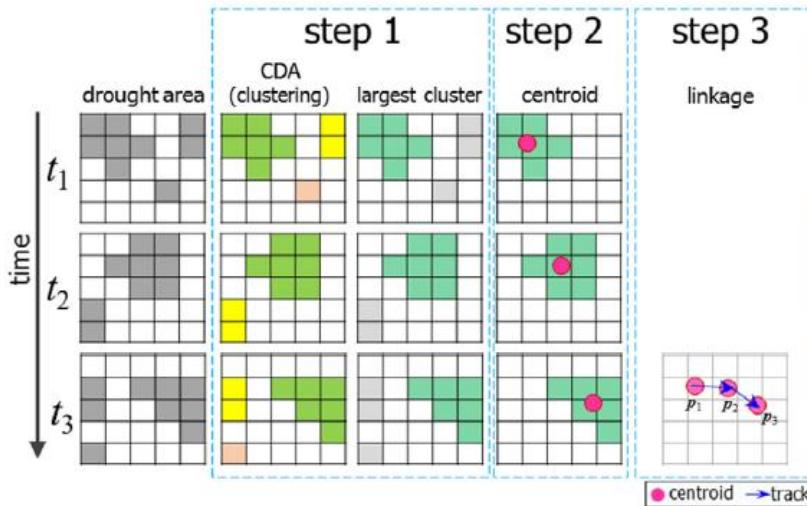


Fig 3. Schematic overview of S-TRACK method for spatial drought tracking which involves: (step 1) spatial drought units (clusters) computation, (step 2) centroids localization, and (step 3) centroids linkage. An example is presented for the case of three time steps: from  $t_1$  to  $t_3$ . Columns in the diagram show the sequence of the steps. Colored cells in the first column indicate all cells in drought. Colors in the second column point out different clusters identified. In the third column, the largest contiguous area in drought is presented with a different color. Only the largest cluster is shown in the fourth column and its centroid ( $p$ ) is indicated by a point. Subscripts indicate time steps.

#### Step 1. Spatial drought unit's computation.

In the spatial context, drought units are identified by means of the Contiguous Drought Area (CDA) analysis (Corzo Perez et al., 2011). These cells in drought are identified in each time step. When the drought indicator is below or equal to the selected threshold, the value of 1 is used to indicate that the cell is in drought, otherwise, the value of 0 is used, indicating non-drought.

Drought indicators (DIs) are representations of a water anomaly. In each time step, the CDAs are computed. The use of CDA relies on the assumption that the binary description of drought condition (0 s and 1 s) is homogeneous over the whole grid. Thus, if two or more cells denote drought conditions (value of 1), and are contiguous in space, it is assumed that all of them are part of the same drought unit. A standardized drought indicator is applied to allows the clustering of neighboring cells in drought (cells with 1 s). After clusters (areas in drought) are identified, the major (largest) one is identified in each time step  $t$  (Fig. 3). As the tracking algorithm focuses on the calculation of the major spatial drought extent in each time step, small or one-cell units are discriminated with the selection of the largest one, allowing the elimination of possible artefact drought areas.

#### Step 2. Centroids localization

After identification of the major (largest) drought cluster, its centroid ( $p$ ) is calculated in each time step. The clusters are joined in time Step 2 and 3 presented are an extension of the CDA analysis. We chose the centroid since we already reduce the spatial representation of drought indicator by using only 1 s and 0 s present for drought and non-drought condition.

#### Step 3. Centroids linkage

The algorithm to link centroids of consecutive clusters in time is a set of rules to separate or join the sequence in time (Fig. 4).

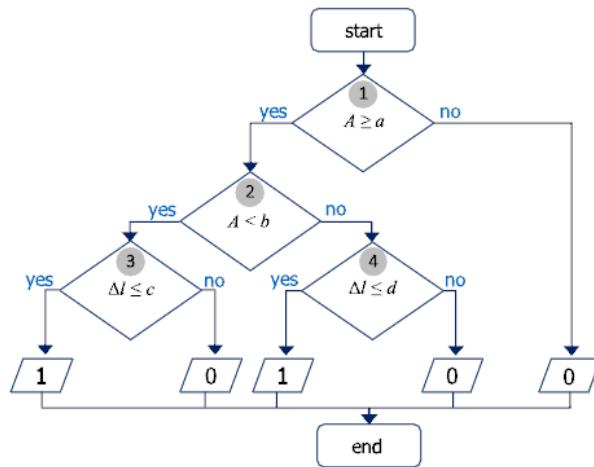


Fig 4. Flowchart showing the rules for linking drought areas (clusters) in time. Numbers in the boxes indicate the sequence of rules 1 to 4. The output of 1 is used to point out that the drought area  $A$  at time  $t$  joins its predecessor at time  $t - 1$ , otherwise 0 is retrieved. The distance between the centroids at times  $t$  and  $t - 1$  is represented by  $\Delta l$ . The linking algorithm has the following

parameters:  $a$ ,  $b$ ,  $c$  and  $d$ . The first two used to control drought area  $A$ , and the last two, to check distance  $\Delta l$ .

The rules consider two types of threshold parameters: (1) two that control the magnitude (size) of the cluster ( $A$ , with dimensions  $L \times L$ ), and (2) two that constrain the Euclidean distance between consecutive clusters ( $\Delta l$ , with dimensions  $L$ ) (Fig. 4). The parameters are denoted as follows:  $a$ ,  $b$ ,  $c$  and  $d$ . The first two are used to the drought area  $A$  and the last two to the distance  $\Delta l$ . The output in this step is the time series of 0 s and 1 s, denoted by  $S(t)$ . Here, the value of 1 indicates the linkage of clusters in time. If the cluster at time  $t$  is not connected with the cluster at time  $t-1$ , the value of 0 is used instead. Consecutive values of 1 s in the time series  $S$  show the occurrence of what is defined as a drought track. The flowchart of the rules for linking the centroids is presented in Fig. 4, and below these rules are explained. Centroids linkage starts by identifying if the cluster area  $A$  is higher than  $a$  (Fig. 4, rule 1). This first comparison helps to discriminate small clusters. If  $A$  is below  $a$ , there is no connection between consecutive clusters and this procedure finalizes, retrieving 0. Before comparing the distance between areas ( $\Delta l$ ), the second comparison of  $A$  is applied to identify if it is a “very large” area (Fig. 4, rule 2). Parameter  $b$  is proposed to consider these large areas. When  $A$  is below  $b$ , the parameter  $c$  is used to compare distances between clusters (Fig. 4, rule 3). Otherwise, when  $A$  is above  $b$  (“very large” area), to restrict the distances, parameter  $d$  is considered instead (Fig. 4, rule 4). The reason of the second comparison of cluster areas and the use of parameter  $d$  is because centroids of clusters with a considerable size may be located farther away from each other and then the distance  $\Delta l$  could fall outside of the limit indicated by parameter  $c$  (Fig. 5).

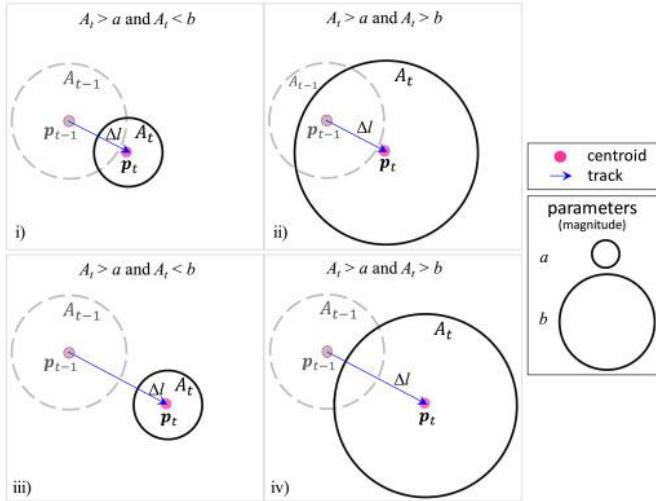


Fig 5. Schematic overview of the four cases of linking clusters (drought areas) in time.

Area at time  $t$  is indicated by  $A_t$  (bold circle) and its predecessor at time  $t-1$  by  $A_{t-1}$  (dashed circle). Centroids of areas  $A_t$  and  $A_{t-1}$  are denoted by  $p_t$  and  $p_{t-1}$  (points), respectively.

Distance between centroids is represented by  $\Delta l$  (arrow). An example of the size of parameters a and b is represented by the circles shown on the right. Centroids in both (i) and (ii) have the same location, in the same way, the centroids in both (iii) and (iv). Areas  $A_t$  in (i) and (iii) are of similar size and between the parameters a and b. On the other hand, in (ii) and (iv), areas  $A_t$  are also equal but above those parameters (case of a “very large” area). Only the parameters of drought area are represented in this figure. Schemes (i) to (iv) help to illustrate the relevance of using parameters that consider not only the magnitude of areas but also the distance between them within the linking algorithm. As a distance limit that helps in linking large areas may not be adequate in connecting smaller ones, as shown in (iv) and (iii), the two distances parameters are proposed in the linking algorithm.

Calculation of drought characteristics to build drought tracks allows for the identification of paths with an onset and an end location. It is possible to extract information regarding the duration, severity, as well as rotation. The proposed approach is called DDRASIC-spatial (Drought duration, severity and intensity computing-spatial events). DDRASIC-spatial is applied after drought tracks are identified by the S-TRACK algorithm. This approach has as a predecessor (Diaz et al., 2019), which however does not consider the elements related to the spatial domain, such as clusters, locations and paths.

For the calculation of the drought duration, firstly the onset and the end are obtained: the time series  $S(t)$  of 1 s and 0 s calculated with S-TRACK method is analyzed to do so. The consecutive sequence of 1 s in the time series  $S$ , indicates the occurrence of a drought track. One isolated value of 1 shows the linking of two clusters in time. Two consecutive values of 1 show the linkage of three clusters in time, and so on. In a sequence of 1s, the time of the first value of 1 ( $t_{first}$ ) is the time step at which the second and first cluster are connected. The time step of the last value of 1 ( $t_{last}$ ) is the one when the last and the penultimate clusters are linked. The onset  $t_i$  is defined as  $t_i = t_{first} - 1$ , while the end  $t_f$  as  $t_f = t_{last}$ . The duration (dd) is calculated follow:

$$dd = \sum_{t=t_i}^{t_f} S(t) \quad (1)$$

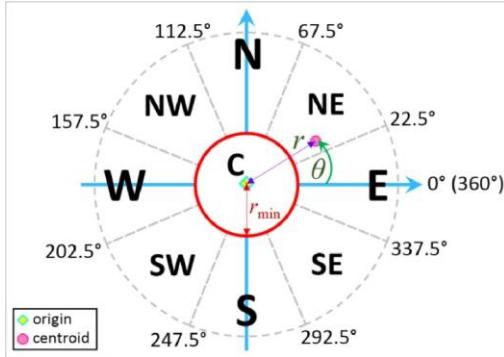
The magnitudes of areas of the largest clusters calculated in each time step with S-TRACK method are saved in the time series DA (drought area). The drought area is used as the measure of the drought severity (ds), which is computed as the sum of drought areas of the period defined by the onset ( $t_i$ ) and the end ( $t_f$ ).

$$ds = \sum_{t=t_i}^{t_f} DA(t) \quad (2)$$

Drought intensity (di) is defined as the ratio between drought severity and duration

$$di = \frac{ds}{dd} \quad (3)$$

Identification of locations where a drought path starts, and ends can provide its main direction. The initial and final locations are identified using the centroids of the first and last cluster, respectively. The location is a relative position in the spatial domain of the study region. It refers to a point in the axes south-north (S-N) and west-east (W-E) (Fig. 6).



*Fig 6.*Schematic overview of the procedure to define centroid's location of a cluster. A centroid can be located in one of nine positions: center (C), east (E), northeast (NE), north (N), northwest (NW), west (W), southwest (SW), south (S) and southeast (SE). The symbol  $r$  stands for the distance between the cluster's centroid and the one of the regions. The angle between the W-E axis and the line defined by centroid's cluster is indicated by  $\theta$ . The radius to define if a cluster is in the center (C) of the region is pointed out by  $r_{\min}$

The origin of the axes is assigned arbitrarily, here it is proposed to place this origin in the centroid of the study region. The centroid of a particular cluster can be located in one of the nine proposed positions: center (C), east (E), northeast (NE), north (N), northwest (NW), west (W), southwest (SW), south (S), and southeast (SE) (Fig. 6). Centre (C) is situated in the centroid of the study region (Fig. 6). A point (centroid) is in the center if the distance ( $r$ ) between such point and the origin is within the  $r_{\min}$  radius (Fig. 6). If distance  $r$  is out of the  $r_{\min}$  radius, the location is assigned based on the angle  $\theta$ . This angle is calculated between the W-E axis and the line defined between the centroid and origin (Fig. 6). Drought tracks provide the visual overview of how drought moves in the spatial domain. Initial and end location (initial and end point of the track) help to identify the direction followed by a given drought cluster. An initial step towards are used to identify and interpret the drought rotation. Drought track can switch between clockwise and counter-clockwise along the pathway, we propose to classify the rotation in a more general way as (1) mostly clockwise (cw), or (2) mostly counter-clockwise (ccw) (Fig. 7).

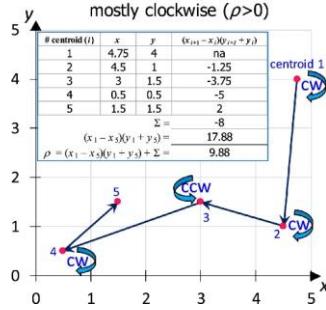
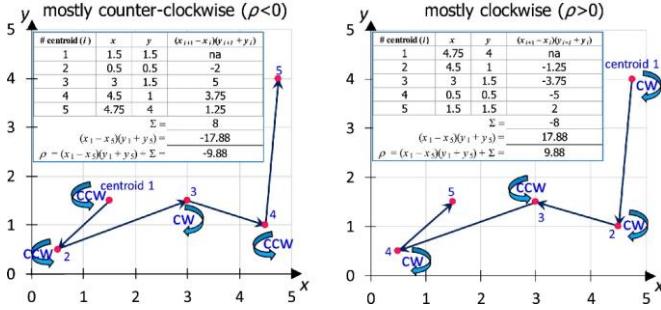


Fig 7. Example of rotation calculation. Two types are considered: (1) mostly counterclockwise when  $\rho < 0$  (left); and (2) mostly clockwise when  $\rho > 0$  (right). The number in each centroid (point) indicates the tracking sequence. Arrows show the track direction and the rotation. Rotation of each line segment is also pointed out by cw and ccw that stand for clockwise and counterclockwise, respectively.

To determine the rotation, a procedure is suggested which makes use of the centroids' coordinates. The algorithm is based on computing a polygon's area ( $A$ ) from the vector with the co-ordinates  $x$  and  $y$  representing the vertices

$$A = \frac{1}{2} |\rho| \quad (4)$$

In this algorithm, firstly the sum of products between the coordinates  $x$  and  $y$ , denoted by  $\rho$  (Eq. (5)), is calculated.

$$\rho = (x_1 - x_n)(y_1 + y_n) + \sum_{i=1}^{n-1} (x_{i+1} - x_i)(y_{i+1} + y_i) \quad (5)$$

Then,  $\rho$  is applied to define the rotation direction (Eq. (6)).

$$\omega = \begin{cases} \text{cw(mostly clockwise) if } \rho > 0 \\ \text{ccw(mostly counter-clockwise) if } \rho < 0 \\ \text{NaN(not defined) if } \rho = 0 \end{cases} \quad (6)$$

The coordinates  $x$  and  $y$  are taken from the ones of centroids' clusters. When there are only two points (two clusters), or when the track is horizontal or vertical, the rotation is not defined, because  $\rho$  takes the value of zero.

### 3. Conclusion

The objective of next study is clear that we should improve spatial extreme drought assessment. However, objective of using GBD, SBD and ML to improve spatial extreme drought extension is

still not transparency. We confuse that should we expect multiple objectives in one study: (i) GBD and SBD combination; (ii) spatial tracking extreme drought. The advantages of combination various methods that it could be efficient to improve results. But in that case, we difficulty go into deep and narrow. Therefore, we suggest separating into various studies: combination GBD and SBD; spatial tracking extreme drought. In previous study, authors figured out some limitations of their research. The clustering only considered drought indicator. Other aspects such as topology, land used were not analysis. Using single meteorological drought index (SPEI) could not fully respect drought phenomena. Based on recommends of authors, we proposal using NDI and considered land cover, DEM in clustering algorithm for tracking prediction spatial extreme drought. If our proposal is accepted, we are going to learn S-TRACK methodology and retrieved DEM, land used data.

## References

- Corzo Perez, G., Van Huijgevoort, M., Voß, F., & Van Lanen, H. (2011). On the spatio-temporal analysis of hydrological droughts from global hydrological models. *Hydrology and Earth System Sciences*, 15(9), 2963-2978.
- Diaz, V., Corzo, G., Van Lanen, H. A., & Solomatine, D. P. (2019). Spatiotemporal Drought Analysis at Country Scale Through the Application of the STAND Toolbox. In *Spatiotemporal Analysis of Extreme Hydrological Events* (pp. 77-93): Elsevier.
- Diaz, V., Corzo, G., Van Lanen, H. A. J., & Solomatine, D. P. (2019). 4 - Spatiotemporal Drought Analysis at Country Scale Through the Application of the STAND Toolbox. In G. Corzo & E. A. Varouchakis (Eds.), *Spatiotemporal Analysis of Extreme Hydrological Events* (pp. 77-93): Elsevier.
- Diaz, V., Corzo Perez, G. A., Van Lanen, H. A. J., Solomatine, D., & Varouchakis, E. A. (2020). An approach to characterise spatio-temporal drought dynamics. *Advances in Water Resources*, 137, 103512. doi:<https://doi.org/10.1016/j.advwatres.2020.103512>
- Meyer, H., Schmidt, J., Detsch, F., & Nauss, T. (2019). Hourly gridded air temperatures of South Africa derived from MSG SEVIRI. *International Journal of Applied Earth Observation and Geoinformation*, 78, 261-267. doi:<https://doi.org/10.1016/j.jag.2019.02.006>
- Mishra, A. K., & Desai, V. R. (2005). Spatial and temporal drought analysis in the Kansabati river basin, India. *International Journal of River Basin Management*, 3(1), 31-41. doi:10.1080/15715124.2005.9635243
- Ren, F.-M., Trewin, B., Brunet, M., Dushmanta, P., Walter, A., Baddour, O., & Korber, M. (2018). A research progress review on regional extreme events. *Advances in Climate Change Research*, 9(3), 161-169. doi:<https://doi.org/10.1016/j.accre.2018.08.001>
- Santos, J. F., Pulido-Calvo, I., & Portela, M. M. (2010). Spatial and temporal variability of droughts in Portugal. *Water resources research*, 46(3). doi:10.1029/2009wr008071

**Professor comments:**

- Make the flowchart for understand what we will do

## WEEK 11

### Spatial extreme drought assessment using natural drought index, satellite-based data and machine learning

#### 1. Materials and Methods

The proposed framework has 5 main steps and is presented in Figure 1. Step 1: collect meteorological data and perform hydrological analysis. Step 2: extract data from the hydrologic model and use observations to construct the input data to estimate the NDI. Step 3: estimate the Natural Drought Index (NDI). Step 4: analyze the spatial tracking extreme drought considered spatial climate pattern such as satellite-based precipitation (SBP). Approximate time series grid cells of NDI, SBP. Classify and Clustering NDI, SBP. Determine link drought cluster centroids. Step 5: analyze the drought duration, drought severity, drought area, drought direction and drought propagation speed.

##### 1.1 Gauge-based precipitation and Satellite based precipitation

The study area covers South Korea has a latitude ( $33^{\circ}$ - $39^{\circ}$  N) and longitude ( $125^{\circ}$ - $131^{\circ}$  E). Data from the Automated Synoptic Observing System (ASOS) network data spanned more than 30 years (1981-2016). It was used as gauge-based precipitation. The ASOS data have an interstation spatial resolution of approximately 12 km.

The spatial extreme drought was assessed from 2014-2015 that was recorded one of the most severe drought periods. For clustering, we used data retrieved from satellites. Monthly precipitation was retrieved from Integrated Multi-satellitE Retrievals (IMERG). The IMERG products include early multi-satellite, late multi-satellite, and final satellite-gauge products with spatial and temporal resolutions of  $1/10^{\circ}$  and 30 min. It is using the unified U.S.-developed algorithm that provides the day-1 multi-satellite precipitation product for the U.S (Huffman et al., 2015). Figure 2 presents a satellite-based precipitation in January 2015 and the locations of the 59 ASOS stations. ASOS provided 9 observed climate variables (daily precipitation, maximum temperature, minimum temperature, average air temperature, dew point temperature, average wind speed, maximum wind speed, average humidity and minimum humidity).

**Commented [VQT2]:** Chỗ này chưa cần trình bày ảnh vệ tinh. Di chuyển xuống dưới

##### 1.2 Variable Infiltration Capacity Model

The variable infiltration capacity (VIC) model has been widely applied for drought monitoring and assessment (Liang et al. 1994). The input data are topography and forcing data. For VIC simulation, the topography includes the terrain and digital elevation model (DEM) provided by the Ministry of Land, Infrastructure, and Transportation (MOLIT). The forcing data include historical observed meteorological data from ASOS. The daily historical runoff was provided by the water resources management information systems and was used to calibrate, verify the VIC model. The model parameters were optimized and verified using observed discharge data from dam inflow sites. The VIC model was simulated with a 24-hour temporal scale and a spatial resolution of  $1/8^{\circ}$  ~ 12.5 km, corresponding to 586 grid cells. Each grid cell simulates the energy flux considering that water and energy balance are performed independently. The topographic and meteorological grids were adjusted to the spatial resolution of the VIC model using the bilinear interpolation

method. Detailed information about the VIC model used in this study can be found in previous studies (Bae et al. 2017).

### **1.3 Natural Drought Index**

Kim et al. (2016) proposed a method to calculate the NDI using the VIC model and principal component analysis (PCA). PCA is a method that produces new variables through a linear combination of the original variables. High-dimensional data can be reduced to low-dimensional data. Precipitation, runoff and soil moisture are the three components that were used in the PCA to compute the NDI. Runoff and soil moisture data were extracted from the VIC model. The infiltration, evapotranspiration and other hydrological processes were included to simulate runoff and soil moisture. These data have a spatial resolution of 1/8 degrees and a temporal scale of 24 hours. Daily precipitation was measured by ASOS at a spatial resolution of 12 km. NDI was analyzed based on the spatial resolution of the 59 ASOS stations and a monthly temporal scale. Precipitation data were matched to the closest runoff and soil moisture grid cell using the nearest-neighbor method. The data were accumulated to match the monthly temporal scale used to compute the NDI. The seasonal influences of variables on drought were also considered; the months were treated separately. An extreme drought event was identified when  $NDI \leq -2$ , which indicated a probability value less than or equal to 0.02 and corresponded to a 50-year return period. The detailed information on the classification of drought based on the NDI can be found from Kim et al. (2016).

### **1.4 Spatial tracking extreme drought**

From 2014-2016 we analyzed 144 images correspond to 36 months of 2 types of data (NDI, SBP). One image has 3338 pixels include study area and surroundings. All the grid cells are inside of study area are marked with code position equal 1, while others are marked with code position equal 0.

NDI, SBD is classified into various levels for clustering. NDI, SBD are divided into 2 levels: extreme drought has a value of 1 if values are below or equal to the 1<sup>st</sup> quartile, and non-extreme drought has a value of 0 if values are higher than the 1<sup>st</sup> quartile.

$$X = \begin{cases} 1 & (X \leq \bar{X} - \sigma) \\ 0 & (X > \bar{X} - \sigma) \end{cases} \quad (1)$$

Where  $\bar{X}$  denote for average of satellite-based precipitation.  $\sigma$  denotes for standard deviation.

We used multivariate clustering analysis (MCA) with K mean algorithm (Jain, 2010; Sajjad et al., 2020). MCA was used to catalog the extreme drought into statistically distinct spatial groups. It is based on the unsupervised machine-learning algorithms and identifies natural clusters in data. Cluster based on attribution of 2 field: NDI (level: 0 or level :1) and SBP (levels:0, 1). Clustering

divide into extreme drought (1) and non-extreme drought (**Figure 3**). Number k of clusters to use is determined by maximizing the Calinski-Harabsz pseudo-F test (Harabasz and Karoński, 1974):

$$Pseudo - F = \left( \frac{A}{W} \right) \left[ \frac{n-k}{k-1} \right] \quad (2)$$

Where A and W are the among-and within-cluster variances, n is the number of objects, and k is the number of existing clusters. Pseudo-F statistics was used to optimize the numbers of cluster with different configuration.

At each time step, we keep only 1 cluster has the largest area of each type of cluster. The distance (d), direction drought clusters ( $\alpha$ ) and speed are determined follow cluster centroid ( $x_t, y_t$ ) and cluster centroid in next step ( $x_{t+1}, y_{t+1}$ ). The speed equals the distance per month.

$$d = \sqrt{(x_t - x_{t+1})^2 + (y_t - y_{t+1})^2} \quad (3)$$

$$\tan \alpha = \frac{|y_t - y_{t+1}|}{d} \quad (4)$$

At each monthly step, we computed drought severity (S) as average NDI of the cluster, and drought area (A) as the total area of that cluster.

$$S = \frac{\sum_{i=1}^n NDI_i}{n} \quad (5)$$

$$A = \sum_i^n a_i \quad (6)$$

Where n is the number of grid cells of this cluster,  $NDI_i$ ,  $a_i$  denoted for NDI value and area of grid cell number i of this cluster.

## 2. Discussion

We are afraid of the duplicated methodology, because step 1 to step 3 was done in our previous study. Should we update the period study to 2019 to avoid it?

**Commented [VQT3]:** Không sợ trùng lắp

-The multivariate clustering analysis, unsupervised machine learning should consider various fields: land cover, topology, satellite-based precipitation or we should consider spatial climate (SBP) as a typology? In NDI various characters such as land cover, meteorology has been considered. The main purpose of our study to obtain spatial extreme drought with support satellite data and machine learning. We propose analysis NDI and considered the impacts of SBP. Because SBP is better explain the spatial characteristics compare to gauge-based precipitation. The other impacts should be considered in the other studies.

-Classify or normalize SBP, NDI is more suitable for our study? Classification is better for clustering, but normalization is better variant of spatial analysis (Figure 4).

-Visualize the clustering result in administrative or grid cells is better? Should we consider administrative boundary in the clustering. It makes the readers easy to follow, but spatial drought does not always follow the administrative boundary (Figure 5). We propose clustering based on grid cells (Figure 5b)

- Although the objective of our study and methodology can be obtained, we are afraid of the interest and attraction. Therefore, we would like to spend more one week to find the solution for the above issues.

Comments from professor:

- direction of study is extreme drought spatial coverages using NDI, SBP, AI

-clear concept before do analysis

## References

- Harabasz, C.T., Karoński, M., 1974. A dendrite method for cluster analysis, Communications in Statistics, pp. 1-27.
- Huffman, G. et al., 2015. Algorithm Theoretical Basis Document (ATBD) Version 4.5: NASA Global Precipitation Measurement (GPM) Integrated Multi-satellitE Retrievals for GPM (IMERG). NASA: Greenbelt, MD, USA.
- Jain, A.K., 2010. Data clustering: 50 years beyond K-means. Pattern Recognition Letters, 31(8): 651-666. DOI:<https://doi.org/10.1016/j.patrec.2009.09.011>
- Sajjad, M., Lin, N., Chan, J.C.L., 2020. Spatial heterogeneities of current and future hurricane flood risk along the U.S. Atlantic and Gulf coasts. Science of The Total Environment, 713: 136704. DOI:<https://doi.org/10.1016/j.scitotenv.2020.136704>

## Appendix

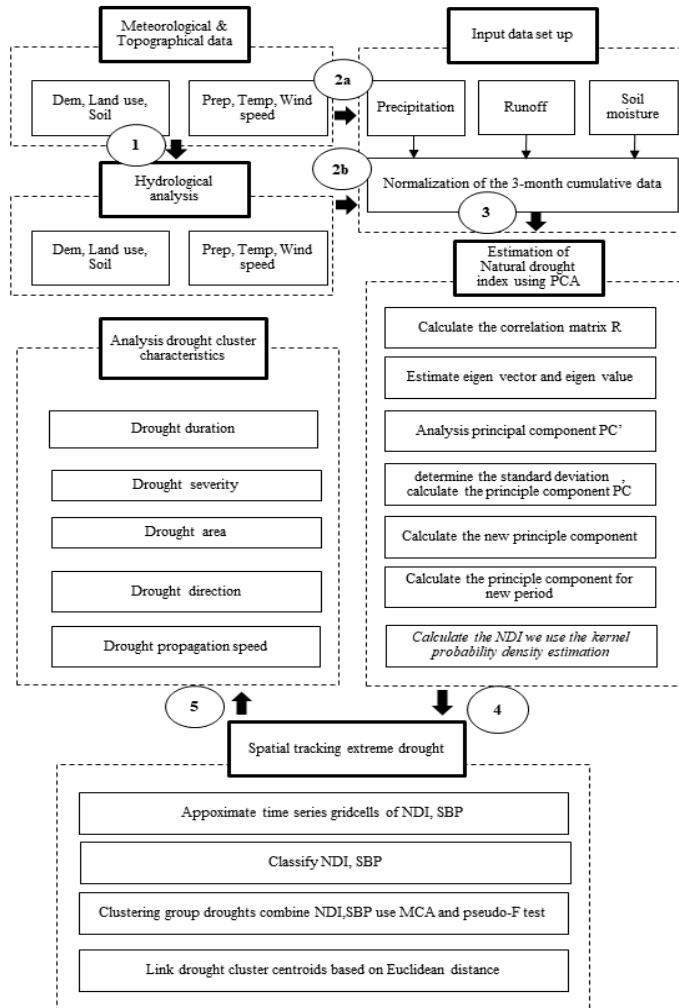


Figure 1. Tracking spatial extreme drought framework

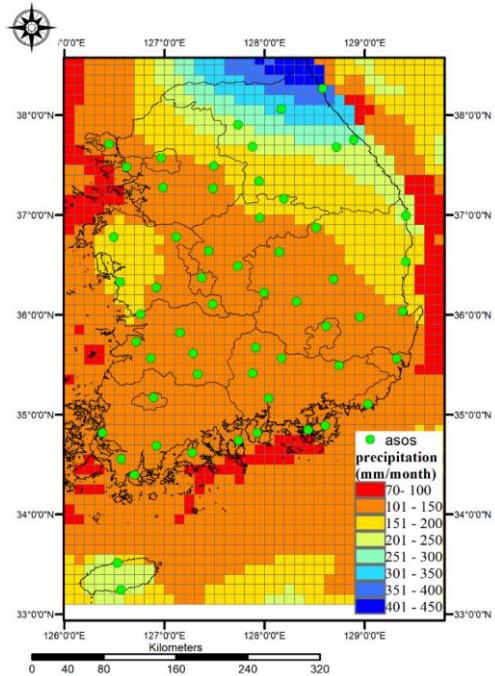


Figure 2. SBP in January and 59 ASOS location

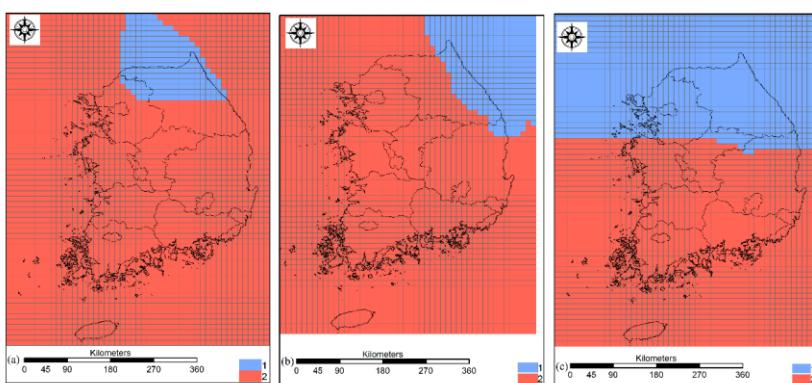


Figure 3. Cluster extreme drought: (a) NDI, (b) SBP, (c) NDI-SBP in January 2015

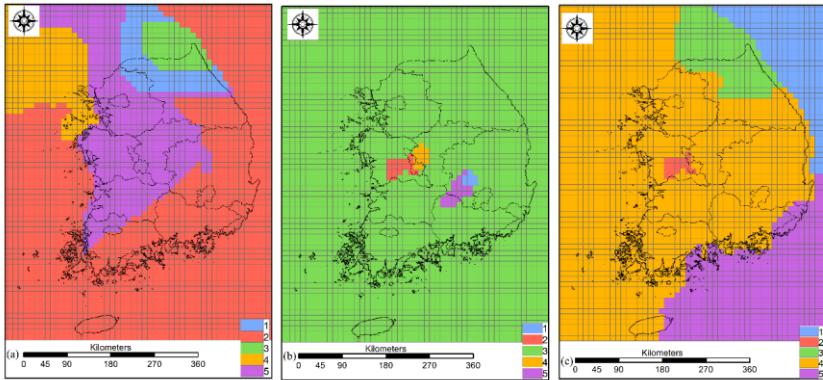


Figure 4. Clustering: (a) GBP, (b) NDI, (c) multivariate GBP and NDI in January 2015

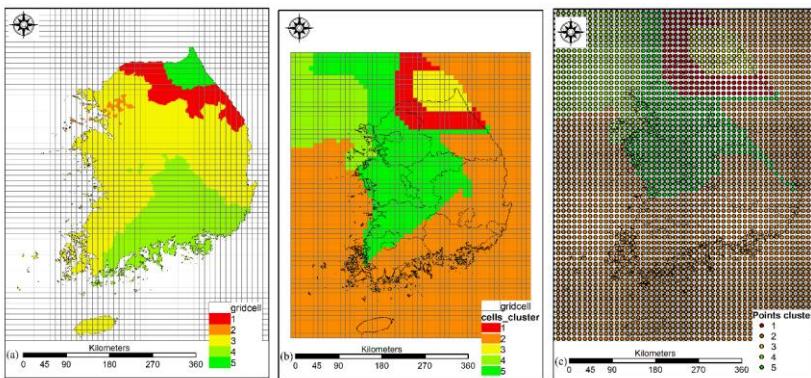


Figure 5. Cluster types: (a) based administration, (b) based on grid cells, (c) based grid points

We proposed multiple clustering to determine extreme drought area

Professor comments:

- Make clear concepts.
- What we will do at each step

## WEEK 12

### Spatial extreme drought assessment using natural drought index, satellite-based data and machine learning

#### 3. Overview the interpolation methods for spatial distribution of extreme drought

Drought is one of the most severe natural disasters. Different with others, it is a creeping phenomenon and occurring in huge region. Therefore, spatial extreme drought analysis is vital to better understand the characteristics of regional drought, and to provide a reference for drought disaster reduction. Several studies have committed to figure out spatial characteristic of extreme drought. The spatial characteristic of drought is usually interpolated from gauge stations data.

Comprehensive comparison of various interpolation was presented in studies (Li and Heap, 2011; Meng et al., 2013). IDW and OK are the most frequently used methods. OK was used to analyze spatial extreme drought for China from 1961-2013 with 810 weather stations (Liu et al., 2016). IDW was used to derive spatial extreme drought for South Korea from 1981-2016 with 59 weather stations (Vo et al., 2020). The spatial extreme drought analysis based on interpolated methods are fast, simple algorithms. However, the accuracy of interpolation depends on the weather stations density that is difficult to be obtained in the reality.

For some cases, intensive observed data obtained from other sources are also available as the auxiliary variables. To utilize the auxiliary information in these data, methods such as regression kriging (RK) or cokriging are proposed. But these methods all assume that the auxiliary variables keep linear correlation with the target variable implicitly, which is not satisfied in most cases. For instance, to improve the interpolation for the region where has limited gauges data, the satellite data is usually utilized to surplus to gauges data. Although of poor quality in value compare to gauges data, the remote sensing data is better on spatial distribution. The combination of satellite-based data and gauge-based data can improve regional drought monitoring and assessment (Bai et al., 2019). Results show that the most suitable gauges density (50-75 gauges per  $10^6 \text{ km}^2$ ) is best for blending and improving the drought monitoring.

In addition, extracting the feature of remote sensing data for merging with gauge-based data, machine learning was utilized. Because it can obtain the complexity spatial pattern of spatial characteristics. Recently, several studies using machine learning approach, satellite data to improve spatial distribution such as digital mapping of soil carbon fraction (Keskin et al., 2019), air temperature (dos Santos, 2020). However, using machine learning, satellite data to improve spatial distribution of extreme drought still less intention.

In this study, we propose using Satellite-based precipitation (SBP) and machine learning (ML) to improve extreme drought spatial distribution (DSD). This study committed to use hybrid model of machine learning and geostatistical approach to improve spatial coverage of extreme drought. The study is organized as follows. Section 2 presents an overview of the materials and methods, including land surface flow modeling, NDI, interpolate method, and hybrid-geostatistical model. Section 3 presents the results and analysis of using the model to study the South Korean region. Section 4 presents some discussion material. Finally, some of the main conclusions about this evaluation model are presented in Section 5.

3.2.

#### **4. Materials and Methods**

The proposed model has 5 main steps and is presented in Figure 1. Step 1: collect meteorological data and perform hydrological analysis. Step 2: extract data from the hydrologic model and use observations to construct the input data to estimate the NDI. Step 3: estimate the Natural Drought Index (NDI). Step 4: Interpolate NDI using IDW and OK. Approximate time series grid cells of NDI, SBP. Building a hybrid regression kriging model (HRK) with SBP as auxiliary variable Step 5: Compare HRK with OK, IDW.

#### **1.5 Gauge-based precipitation and Satellite based precipitation**

The study area covers South Korea has a latitude ( $33^{\circ}$ - $39^{\circ}$  N) and longitude ( $125^{\circ}$ - $131^{\circ}$  E). Data from the Automated Synoptic Observing System (ASOS) network data spanned more than 30 years (1981-2016). It was used as gauge-based precipitation. The ASOS data have an interstation spatial resolution of approximately 12 km. The spatial extreme drought was assessed from 2014-2015 that was recorded one of the most severe drought periods. ASOS provided 9 observed climate variables (daily precipitation, maximum temperature, minimum temperature, average air temperature, dew point temperature, average wind speed, maximum wind speed, average humidity and minimum humidity).

#### **1.6 Variable Infiltration Capacity Model**

The variable infiltration capacity (VIC) model has been widely applied for drought monitoring and assessment (Liang et al. 1994). The input data are topography and forcing data. For VIC simulation, the topography includes the terrain and digital elevation model (DEM) provided by the Ministry of Land, Infrastructure, and Transportation (MOLIT). The forcing data include historical observed meteorological data from ASOS. The daily historical runoff was provided by the water resources management information systems and was used to calibrate, verify the VIC model. The model parameters were optimized and verified using observed discharge data from dam inflow sites. The VIC model was simulated with a 24-hour temporal scale and a spatial resolution of  $1/8^{\circ}$  ~ 12.5 km, corresponding to 586 grid cells. Each grid cell simulates the energy flux considering that water and energy balance are performed independently. The topographic and meteorological grids were adjusted to the spatial resolution of the VIC model using the bilinear interpolation method. Detailed information about the VIC model used in this study can be found in previous studies (Bae et al. 2017).

#### **1.7 Natural Drought Index**

Kim et al. (2016) proposed a method to calculate the NDI using the VIC model and principal component analysis (PCA). PCA is a method that produces new variables through a linear combination of the original variables. High-dimensional data can be reduced to low-dimensional data. Precipitation, runoff and soil moisture are the three components that were used in the PCA to compute the NDI. Runoff and soil moisture data were extracted from the VIC model. The infiltration, evapotranspiration and other hydrological processes were included to simulate runoff

and soil moisture. These data have a spatial resolution of 1/8 degrees and a temporal scale of 24 hours. Daily precipitation was measured by ASOS at a spatial resolution of 12 km. NDI was analyzed based on the spatial resolution of the 59 ASOS stations and a monthly temporal scale. Precipitation data were matched to the closest runoff and soil moisture grid cell using the nearest-neighbor method. The data were accumulated to match the monthly temporal scale used to compute the NDI. The seasonal influences of variables on drought were also considered; the months were treated separately. An extreme drought event was identified when  $NDI \leq -2$ , which indicated a probability value less than or equal to 0.02 and corresponded to a 50-year return period. The detailed information on the classification of drought based on the NDI can be found from Kim et al. (2016).

### 1.8 Geostatistical interpolation

NDI from 59 stations are interpolated with various methods. The general estimation equation of spatial interpolation approach is given as follows:

$$\hat{Z}(x_0) = \sum_{i=1}^N w_i Z(x_i) \quad (1)$$

Where  $\hat{Z}(x_0)$  represents the estimated value of the prediction parameter  $Z$  at the unsampled location  $x_0$ ,  $Z(x_i)$  is the actual value at the interpolating point  $x_i$ ,  $w_i$  is the weight of the interpolating point  $x_i$  and  $N$  is the number of sample point. Spatial interpolation methods divide into 2 classes. The first class is deterministic interpolation methods including Inverse Distance Weighting (IDW), spline, and Radial Basis Functions (RBF). Deterministic interpolation techniques create surfaces from sample points using mathematical functions, based on either the extent of similarity (IDW) or the degree of smoothing (RBF). The simple IDW estimate  $\hat{Z}(x_0)$  follow:

$$\hat{Z}(x_0) = \sum_{i=1}^N \frac{w_i x_i}{\sum_{j=1}^N w_j} \quad (2)$$

Where  $N$  is the number of sampled locations,  $x_i$  is the parameter value at the  $i$ th location, and  $w_i$  is weight of the  $i$ th interpolating point. The assigned weight by IDW is expressed as follows:

$$w_i = \frac{1}{d(x_0, x_i)^p} \quad (3)$$

Here,  $d(x_0, x_i)$  is the distance between the prediction point  $x_0$  and the  $i$ th interpolating point  $x_i$ ,  $p$  is the power factor.

The second class is stochastic methods such as Ordinary Kriging (OK), Universal Kriging (UK) and cokriging (COK). It quantifies the spatial autocorrelation among sampling points and accounts

for the spatial configuration of the sampling points around the prediction location. It represents the family of generalized least square regression-based interpolation methods. It aims to minimize mean squared error in prediction.

$$\hat{Z}(x_0) = \sum_{i=1}^N w_i [Z(x_i) - \mu(x_0)] + \mu \quad (4)$$

Here,  $\hat{Z}(x_0)$  is the predicted parameter value at point  $x_0$ ,  $\mu$  is the constant mean value over the region of interest. The OK assumes the stationarity of first moment of the prediction parameter, that is  $E\{Z(x_i)\} = E\{Z(x_0)\} = \mu = \mu(x_0)$ ,  $\mu$  is unknown. The  $w_i$  is measured from the experimental semivariogram.

$$\gamma(h) = \frac{\sum_{i=1}^M [Z(x_i) - Z(x_i + h)]^2}{2M} \quad (5)$$

Here,  $\gamma(h)$  represents the semivariance at lag interval  $h$ ,  $Z(x_i)$  is measured parameter value at point  $x_i$ ,  $Z(x_i + h)$  is measured parameter value at sampled location which is  $h$  lag distance apart from  $x_i$ ,  $M$  is the total number of pairs of the interpolating points that are  $h$  distance lag apart. Detail of semantic Kriging for spatial-temporal prediction can reference in studies (Bhattacharjee et al., 2019).

The Regression Kriging interpolation (RK) is formally defined by Hengl et al. (2007). It refers to a special kind of the UK algorithm that uses auxiliary variable for the external drift estimation variable. The word drift is term for trends in geo-statistics, and the term external drift is the drift extracted from auxiliary variables.

$$\hat{Z}_{RK}(x_0) = \beta_0 + \sum_{k=1}^K \beta_k y_k(x_0) + \sum_{i=1}^K \lambda_i (z(x_i) - \beta_0 - \sum_{k=1}^K \beta_k y_k(x_i)) \quad (6)$$

Here,  $\hat{Z}_{RK}(x_0)$  represents the corresponding value of target variable,  $\beta_0$  is the constant coefficient,  $\beta_k$  represents the coefficient of the  $k$ th auxiliary variable  $y_k$ ,  $\lambda_i$  is the coefficient for the regression residual of  $x_i$ , while  $\beta_k$  represents the coefficient of the  $k$ th auxiliary variable  $y_k(x_i)$ . In RK,  $\beta = (\beta_0 \dots \beta_K)$  are computed by generalized least squares interactions. The initial coefficients  $\beta^{(0)}$  are obtained by the ordinary least square's regression between  $z$  and  $y_1 \dots y_K$ . Denoting the coefficients after the  $t$ th iteration as  $\beta^{(t)}$ , then for the next iteration

$$\beta^{(t+1)} = (A^T C^{t-1} A)^{-1} \cdot A^T C^{t-1} \cdot z \quad (7)$$

Here,  $A$  is an  $N$  rows  $K+1$  column matrix of the auxiliary variables, which takes the constants vector 1 as the first column, and values of  $y_1 \dots y_K$  as the rest  $K$  columns.  $C^{(t)}$  is the covariance matrix for the residuals  $r^{(t)} = z - A\beta^{(t)}$  obtained after the  $t$ th iteration. The above iterations will keep executing until  $\|\beta^{(t+1)} - \beta^{(t)}\|$  is converged to zero. One  $\beta_0 \dots \beta_K$  are determined, the

regression residuals are able to be computed by  $r(x_i) = z(x_i) - \beta_0 - \sum_{k=1}^K \beta_k y_k(x_i)$ . By statistically fitting the experimental variograms of  $r(x_1) \dots r(x_N)$ , where the variogram function

$$\begin{aligned} & \left\{ \sum_{i=1}^N \lambda_i \cdot \gamma_r(x_i, x_1) = \gamma_r(x_0, x_1) \right. \\ \gamma_r = & \left\{ \sum_{i=1}^N \lambda_i \cdot \gamma_r(x_i, x_1) = \gamma_r(x_0, x_N) \right. \\ & \left. \left| \sum_{i=1}^N \lambda_i = 1 \right. \right. \end{aligned} \quad (8)$$

The RK has intuitive two-phase explanation if it is separated

$$\begin{cases} F_{LR}(x_0) = \beta_0 + \sum_{k=1}^K \beta_k y_k(x_i) \\ F_{SK}(x_0) = \sum_{i=1}^N \lambda_i (z(x_i) - F_{LR}(x_0)) \end{cases} \quad (9)$$

During the first phase, a regression model between  $z$  and  $y_1 \dots y_K$  is established by the iterative generalized least squares regression, which provides a linear fitting  $F_{LR}(x)$  for the drift term  $T_z(x)$ . Then, by the interpolation of simple kriging, the residual term  $R_z(x)$  is estimated as  $F_{LR}(x)$

### 1.9 Hybrid neural network-kriging model

In hybrid neural network-kriging model, the function  $Z_M(y_1, \dots, y_K)$  is equivalent to a nonlinear fitting for the drift term  $T_z(x)$

$$F_{NLR}(x_0) = Z_M(y_1(0), \dots, y_K(x_0)) \quad (10)$$

We replaced the linear fitting term  $F_{LR}$  with nonlinear fitting  $F_{NLR}$ , which is used machine learning regression instead of generative least squares. Finally, we obtain a hybrid model of machine learning and kriging

$$Z_{HRK}(x_0) = \beta_0 + \beta_1 \cdot \hat{y}(x_0) + \sum_{i=1}^N \lambda_i (z(x_i) - \beta_0 - \beta_1 \cdot \hat{y}(x_i)) \quad (11)$$

Data retrieved from satellites are used as auxiliary variable to build the hybrid machine learning-geostatistical model. Monthly precipitation was retrieved from Integrated Multi-satellitE Retrievals (IMERG). The IMERG products include early multi-satellite, late multi-satellite, and final satellite-gauge products with spatial and temporal resolutions of  $1/10^\circ$  and 30 min. It is using the unified U.S.-developed algorithm that provides the day-1 multi-satellite precipitation product for the U.S (Huffman et al., 2015). Figure 2 presents a satellite-based precipitation in January 2015

and the locations of the 59 ASOS stations. From 2014-2016 we analyzed 144 images correspond to 36 months of 2 types of data (NDI, SBP). One image has 3338 pixels include study area and surroundings.

## 5. Discussion

-Compare with multivariate clustering analysis, the hybrid model of machine learning and geostatistical is more suitable for our objective. However, for handling whole the theoretical and application should take a time. Therefore, we propose spending more time to understand more theoretical, and handle data programming. It means we would like to learn more about how to integrate neural network with kriging interpolation methods.

- Our methodology should clarify about how to use neural network replace for generative least square regression. The proposal comparison in step 5 of methodology is simple. Should we find some other metrics more interesting.

## References

- Bai, X., Wu, X., Wang, P., 2019. Blending long-term satellite-based precipitation data with gauge observations for drought monitoring: Considering effects of different gauge densities. *Journal of Hydrology*, 577: 124007. DOI:<https://doi.org/10.1016/j.jhydrol.2019.124007>
- Bhattacharjee, S., Ghosh, S.K., Chen, J., 2019. Semantic Kriging for Spatio-temporal Prediction, 839. Springer.
- dos Santos, R.S., 2020. Estimating spatio-temporal air temperature in London (UK) using machine learning and earth observation satellite data. *International Journal of Applied Earth Observation and Geoinformation*, 88: 102066. DOI:<https://doi.org/10.1016/j.jag.2020.102066>
- Hengl, T., Heuvelink, G.B., Rossiter, D.G., 2007. About regression-kriging: From equations to case studies. *Computers & geosciences*, 33(10): 1301-1315.
- Keskin, H., Grunwald, S., Harris, W.G., 2019. Digital mapping of soil carbon fractions with machine learning. *Geoderma*, 339: 40-58. DOI:<https://doi.org/10.1016/j.geoderma.2018.12.037>
- Li, J., Heap, A.D., 2011. A review of comparative studies of spatial interpolation methods in environmental sciences: Performance and impact factors. *Ecological Informatics*, 6(3): 228-241. DOI:<https://doi.org/10.1016/j.ecoinf.2010.12.003>
- Liu, X.-F. et al., 2016. Spatial analysis of meteorological drought return periods in China using Copulas. *Natural Hazards*, 80(1): 367-388. DOI:10.1007/s11069-015-1972-7
- Meng, Q., Liu, Z., Borders, B.E., 2013. Assessment of regression kriging for spatial interpolation – comparisons of seven GIS interpolation methods. *Cartography and Geographic Information Science*, 40(1): 28-39. DOI:10.1080/15230406.2013.762138
- Vo, Q.-T., So, J.-M., Bae, D.-H., 2020. An Integrated Framework for Extreme Drought Assessments Using the Natural Drought Index, Copula and Gi\* Statistic. *Water Resources Management*. DOI:10.1007/s11269-020-02506-7

## Appendix

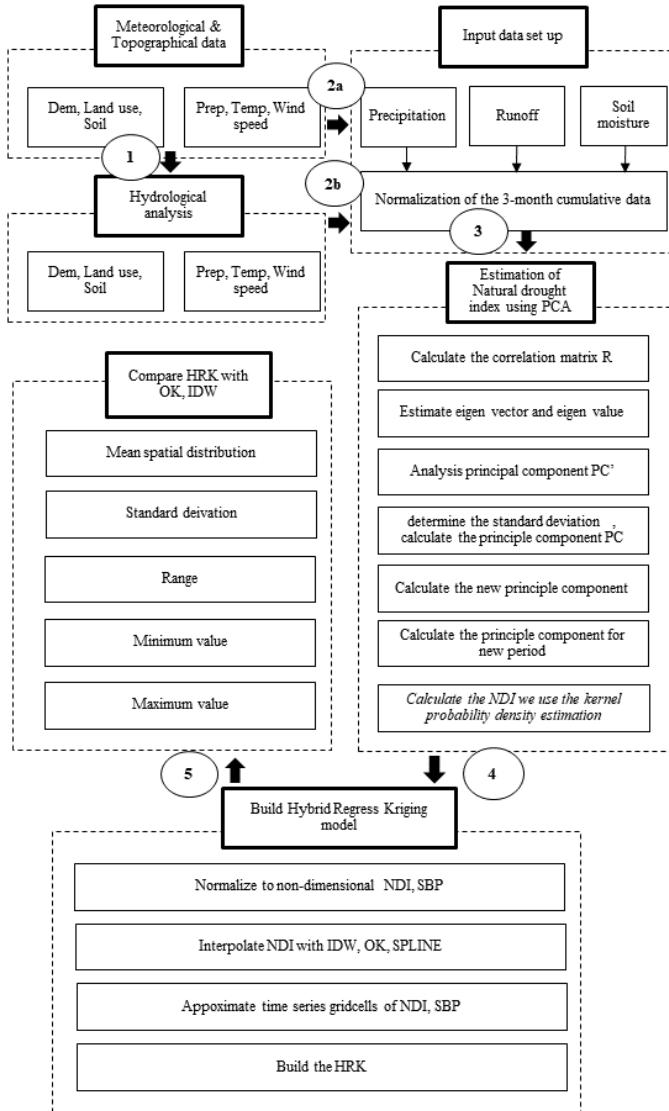


Figure 6. Hybrid Neural network-regression kriging model to estimate spatial extreme drought

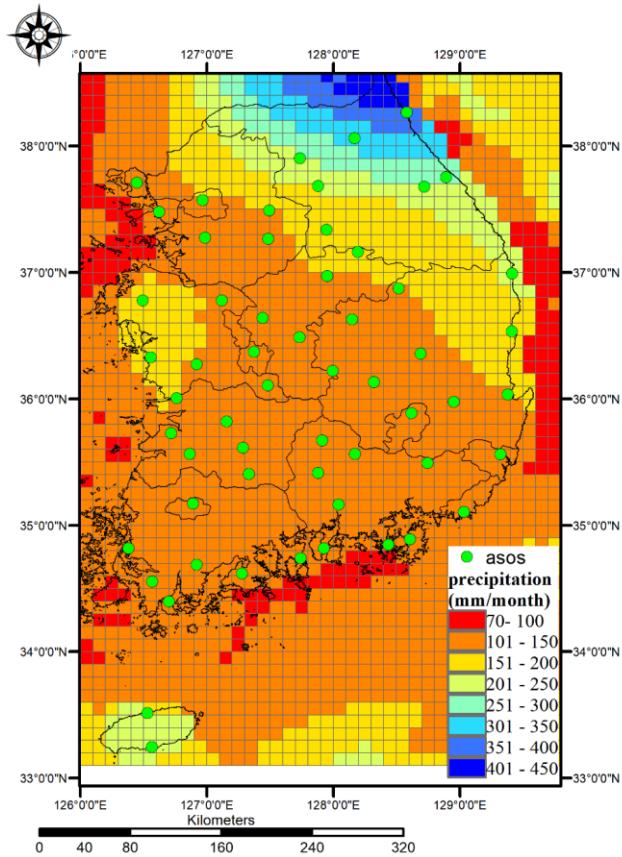


Figure 7. SBP in January 2015 and 59 ASOS location

We propose using hybrid regression Kriging and ANN to prediction spatial extreme drought

Professor agreed to use this concept to improve spatial coverages but need to make it clear

## WEEK 13

### Spatial extreme drought assessment using natural drought index, satellite-based data and hybrid Kriging approach

#### 6. Hybrid Regression model

Spatial extreme drought was predicted by using four predictors: natural drought index (NDI), satellite-based precipitation (SBP), bias of natural drought index & satellite-based precipitation (BSNDISBP), and elevation map. Model was setup follow the processing. First, NDI was interpolated by using Ordinary Kriging (OK). The spatial resolution of NDI was chosen as the same as SBP and elevation map ( $0.1 \text{ degree} \times 0.1 \text{ degree}$ ). We continue computed the bias of NDI and SBP at each grid cells and interpolated by OK to create BSNDISBP grid. The spatial resolution was consistency ( $0.1 \text{ degree} \times 0.1 \text{ degree}$ ). Then, we computed sample induced NDI, SBP, BSNDISBP, and elevation. Here, NDI was the target value, while SBP, BSNDISBP and elevation were auxiliary values. They were considered as predictors.

Second, the sample are randomized by splitting 10 samples using cross validation (CV) to estimate 7 models of the next step. We utilized interpolation, regression, hybrid regression kriging approach. The interpolation approach was used with OK. The regression approach was used GLM (generalized linear model), GAM (generalized additive model), and RF (random forest). While hybrid regressing kriging approach was used GLM+OK, GAM+OK, and RF+OK of residuals.

To evaluate modes, we use 7 indices included Coefficient of efficiency (CE), Coefficient of determination( $r^2$ ), Index of agreement (d), Mean square relative error (MSRE), Mean absolute error (MAE), Root mean square error (RMSE), Mean square error (MSE).

3.3.

#### 7. Initial results

**Figure 1** presents the spatial interpolation of NDI, SBP, BSNDISBP, and elevation. They show the various spatial distribution. NDI has the higher value in middle-west area. SBP has the high value at the bottom east and west area. The bias of the NDI and SBP occurs almost in the center of area where is lower elevation.

**Figure 2** show the normalize NDI of 59 station in August 2015. The most frequency value is 0.5 or  $0.5 \times (2.75 - (-3.25)) + (-3.25) = -0.25$ . Here, maximum value of NDI is 2.75, minimum value of NDI is -3.25. The observation NDI locates almost in the center area.

**Figure 3** show the correlation of various variable (NDI, SBP, BSNDISBP, elevation, longitude X, latitude Y). NDI has the most correlation to BSNDISP (0.92), light correlation to others: longitude (0.36), latitude (0.23), elevation (0.2), SPI (0.25). It shows BSNDISBP is the most important factors to predict spatial drought distribution.

The mean of RMSE of cross validate for 10 samples shows RF and RF\_OK has the best results with mean of RMSE (0.4). While the conventional interpolated method OK has the largest mean RMSE (0.101). It shows the RF and RF\_OK can improve spatial extreme drought coverage.

**Figure 4** shows the spatial drought distribution of RF\_OK.

## 8. Discussion

Based on the initial results, some issues should be considered:

-Spatial resolution 0.1 degree × 0.1 degree approximate 10 km × 10 km is coarse. We propose using some method to upscale resolution to 1 km× 1km. It will show more clearly spatial distribution.

- Should we extent the auxiliary variables such as evapotranspiration, land cover, temperature, from satellite? Because we still don't not know the correlation of NDI with other variables. In data driven approach, we thought more relevant data, can improve more result.
- Should we reduce the performer evaluation metrics. We have seen some studies using a lot metrics like that, but we would like to reduce it to make our study is more concise.
- Almost above methods were based on linearity. We would like to build a simple neural network for prediction in the future. Because Neural network can handle the non-linearity problems. The current methodology (**Figure 5**) will be modified by adding artificial neural network regression kriging (ANNRKR) method.

## Appendix

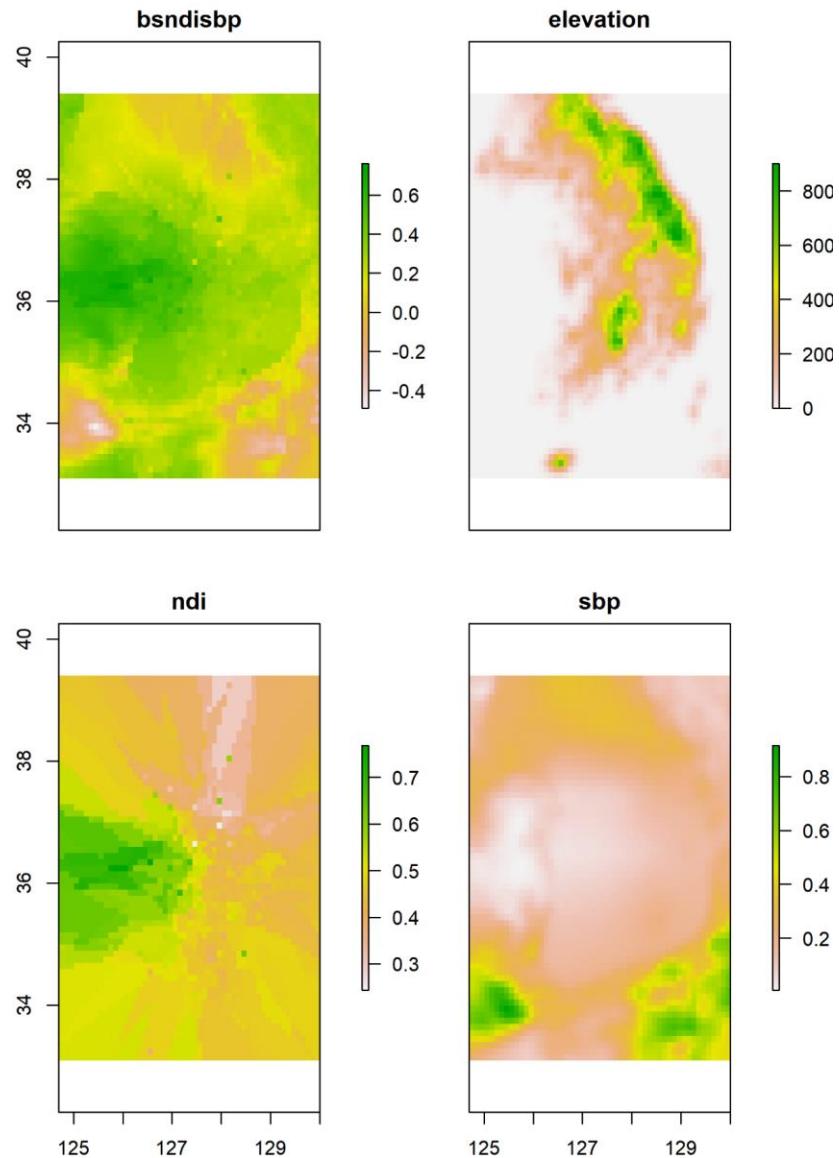


Figure 8. spatial interpolation NDI (natural drought index), satellite-based-precipitation (SBP), bias NDI & SBP, elevation.

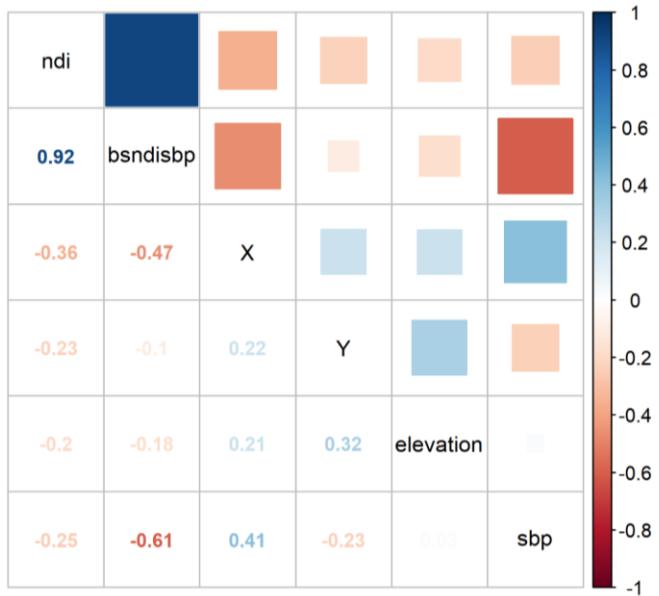


Figure 9. Correlation of various predictors

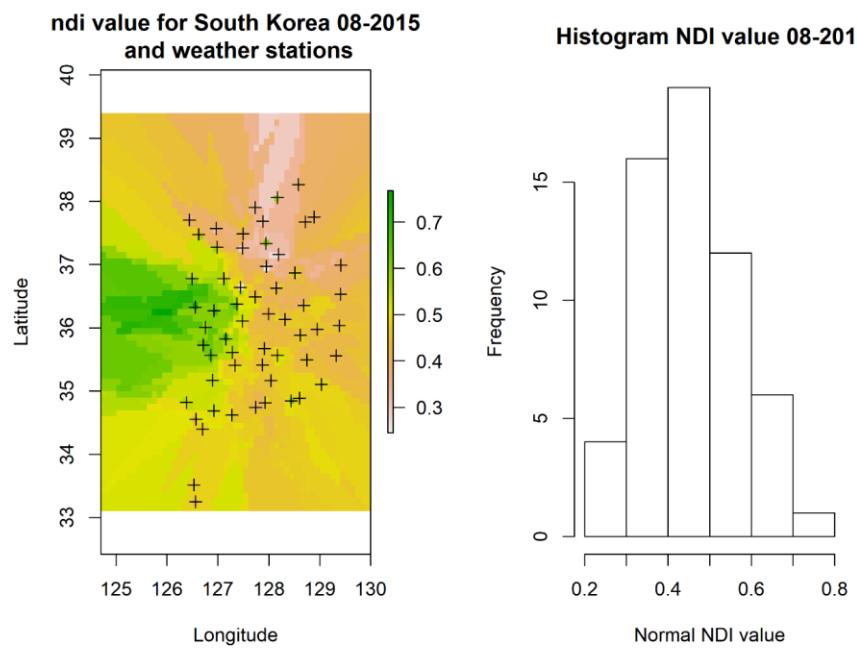


Figure 10. Histogram of NDI in August 2015

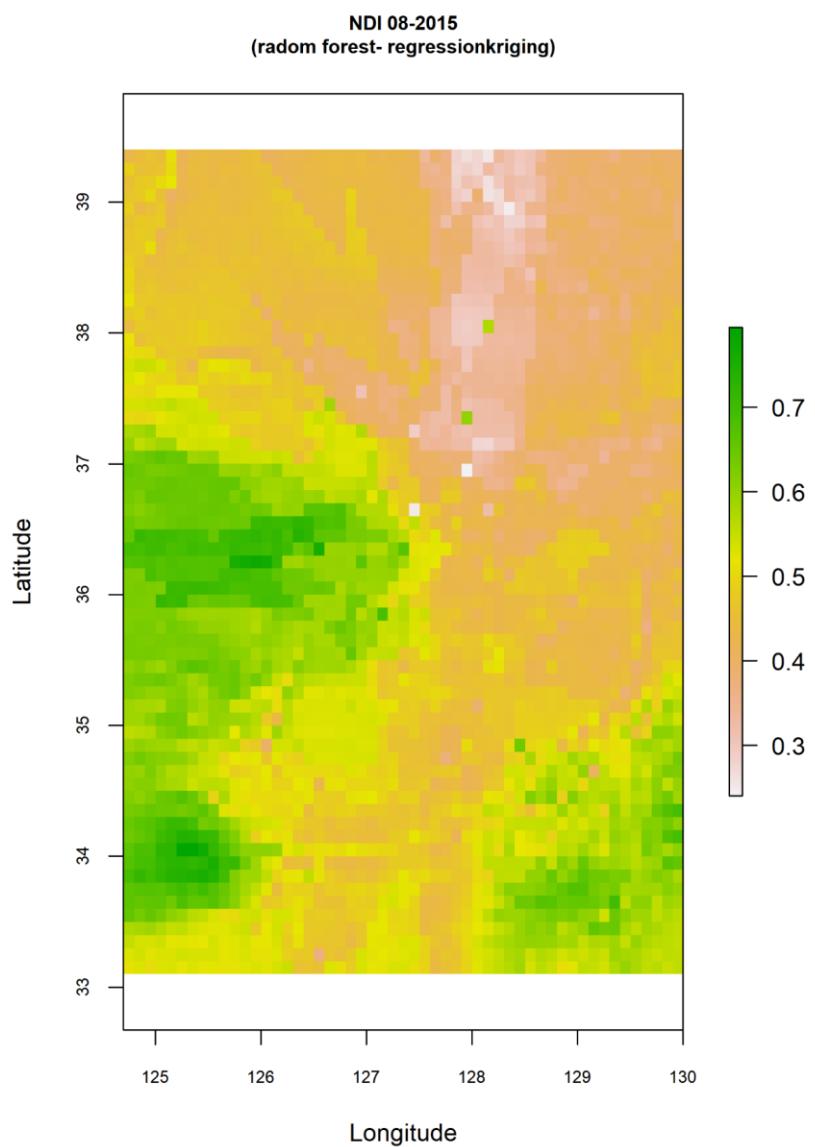


Figure 11. Prediction spatial extreme drought using Random Forest Regression Kriging

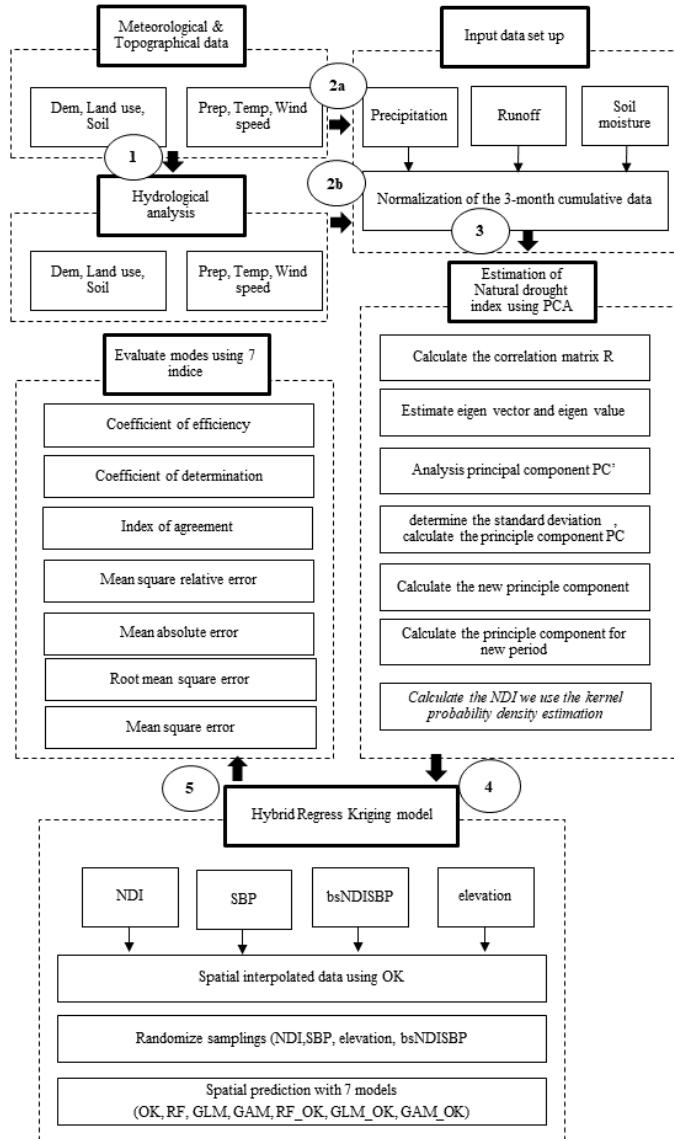


Figure 12. Hybrid egression kriging model to estimate spatial extreme drought



## WEEK 14

### **Spatial extreme drought assessment using natural drought index, satellite-based data and hybrid Kriging approach**

#### **9. Determine number of auxiliary variables**

Our model was extended by adding two auxiliary variables: temperature and soil moisture. Data were retrieved from satellite. They are important in the hydrological process. The higher temperature make evaporation is severe lead to more drought. While soil moisture directly impacts to drought phenomenal. The spatial distribution of data set was present in Figure 1. We analyzed the correlation of them. Temperature and Soil moisture almost in-corelate to NDI. The correlation of NDI ~ temperature (0.0787) and NDI ~ soil moisture (0.0501) were shown in Figure 2.

#### **10. Neural network regression**

We used sample neural network models (1 hidden layer) and deep learning (multiple hidden layers) to setup neural network regression kriging model (NN). Diagram of a sample neural network are shown in Figure 3. In Figure 3 shows a neural network model with 1 hidden layer included 5 epochs (neurons). The number of epochs was selected based on trial. Number of hidden layers was also chosen by trials. In this study we test model from 1 to 3 layers. Data was split into 80% for training and 20% for testing. We used logistic as active function.

Base on RMSE, 1 hidden layers-5 epochs give the lowest RMSE (Figure 4). However, split this sample into 10 random samples to evaluate models, we found that accuracy of NN is only higher performance compare with original kriging (Figure 5). It is less accurate compare to others (RF, GLM, GAM, RF\_OK, GAM\_OK). These models based on linearity approach. It is can be explained by using the non-optimal neural networks algorithms. Neural network model is needed to optimize and turning parameter. because this result is opposite to output of study (Seo et al., 2015). In this study, Neural network regression was overperformance compare to other methods.

On the other hand, we supposed the regression results were depended on data sample. It is difficult to determine the model can fit all data. Therefore, our data were fit better with GLM\_OK.

#### **11. Discussion**

We finished added Neural Network to our framework (Figure 6). However, the results should be improving more. We propose spending more time to optimize parameters of models and fully understand methods to estimate performance of Neural Network. Beside that we would like to review other linearity models for comparison. What are advantages and disadvantages of them? If Neural Network is always better than conventional methods?

#### **References**

Seo, Y., Kim, S., & Singh, V. P. (2015). Estimating spatial precipitation using regression kriging and artificial neural network residual kriging (RKNNRK) hybrid approach. *Water resources management*, 29(7), 2189-2204.

## Appendix

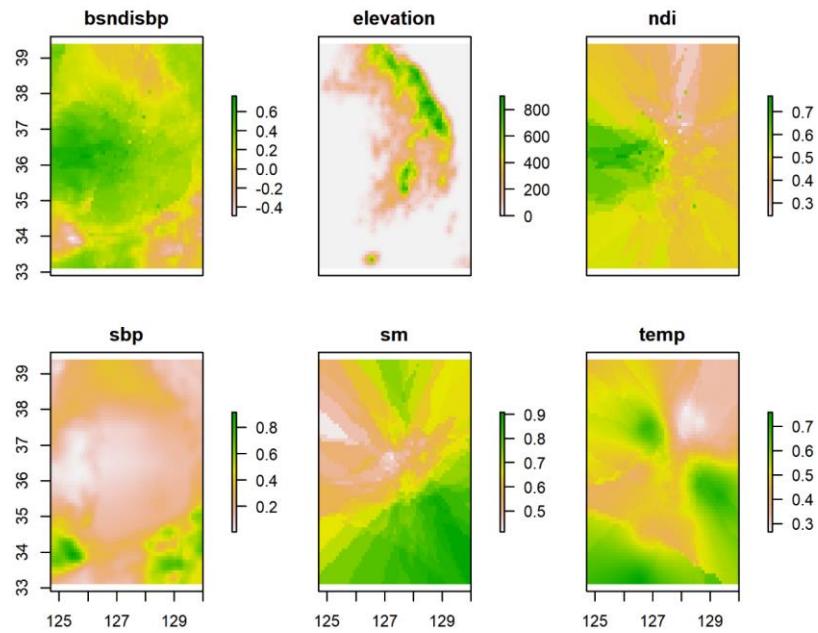


Figure 13. Spatial distribution of dataset

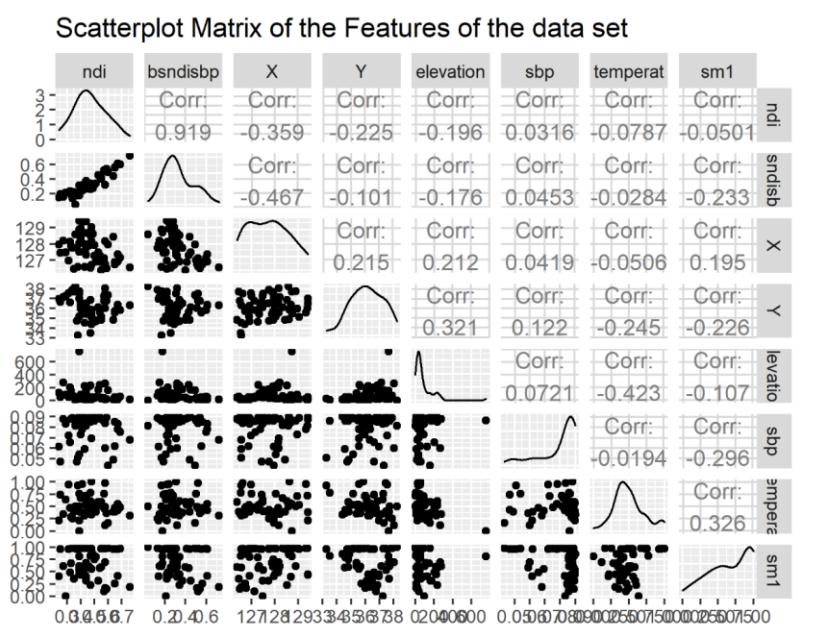


Figure 14. Correlation of data set

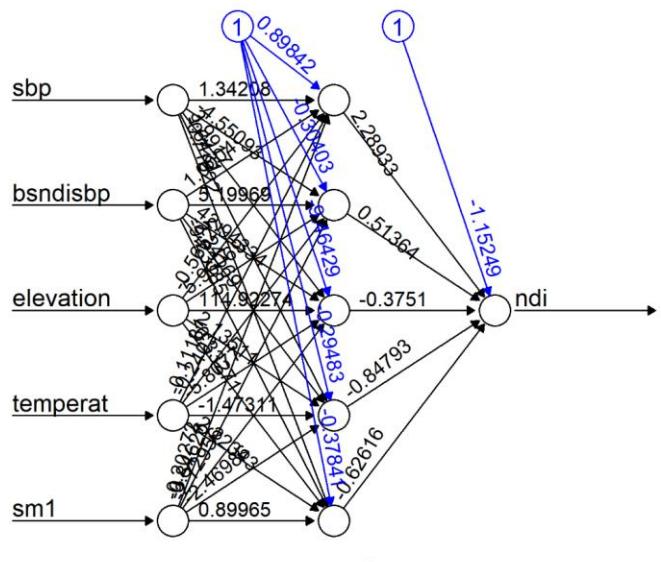


Figure 15. Neural network of model  $NDI \sim SBP + BSNDI + Elevation + Temperature + Soil moisture$

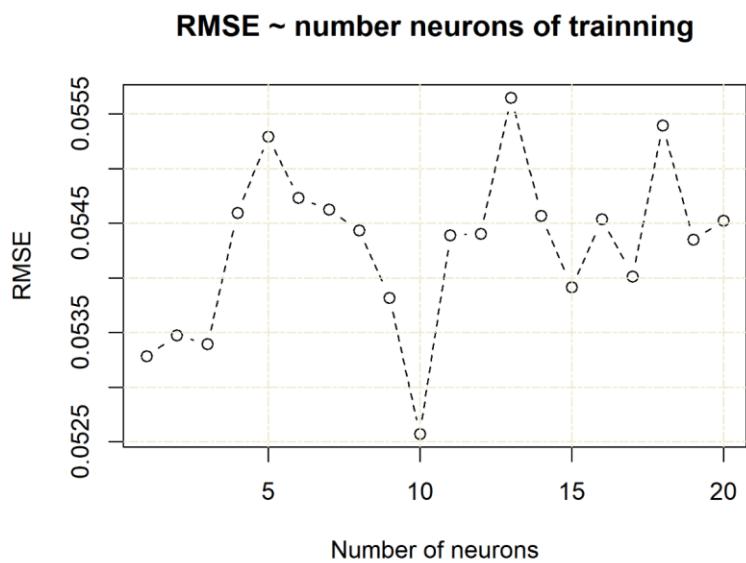


Figure 16. Error of training depend on the number of neuron (epoch)

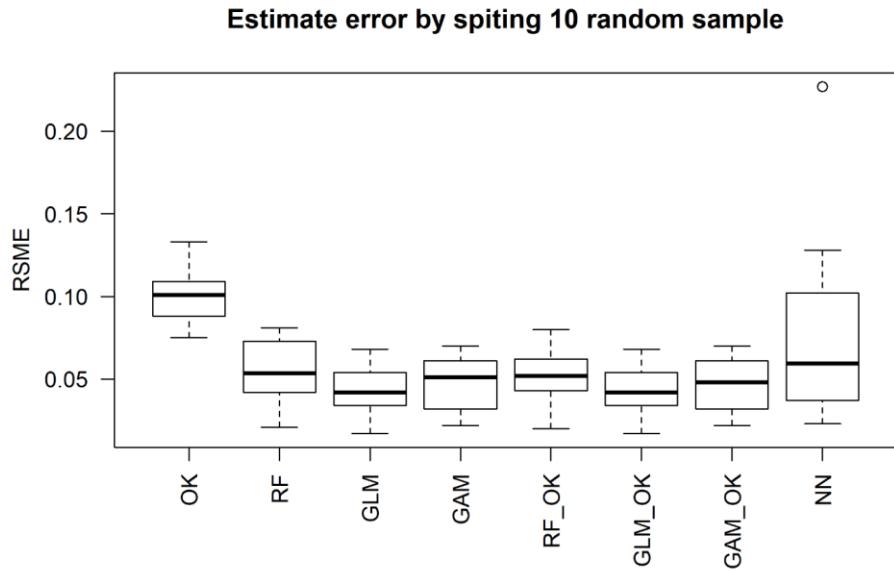


Figure 17. Compare the performance of models

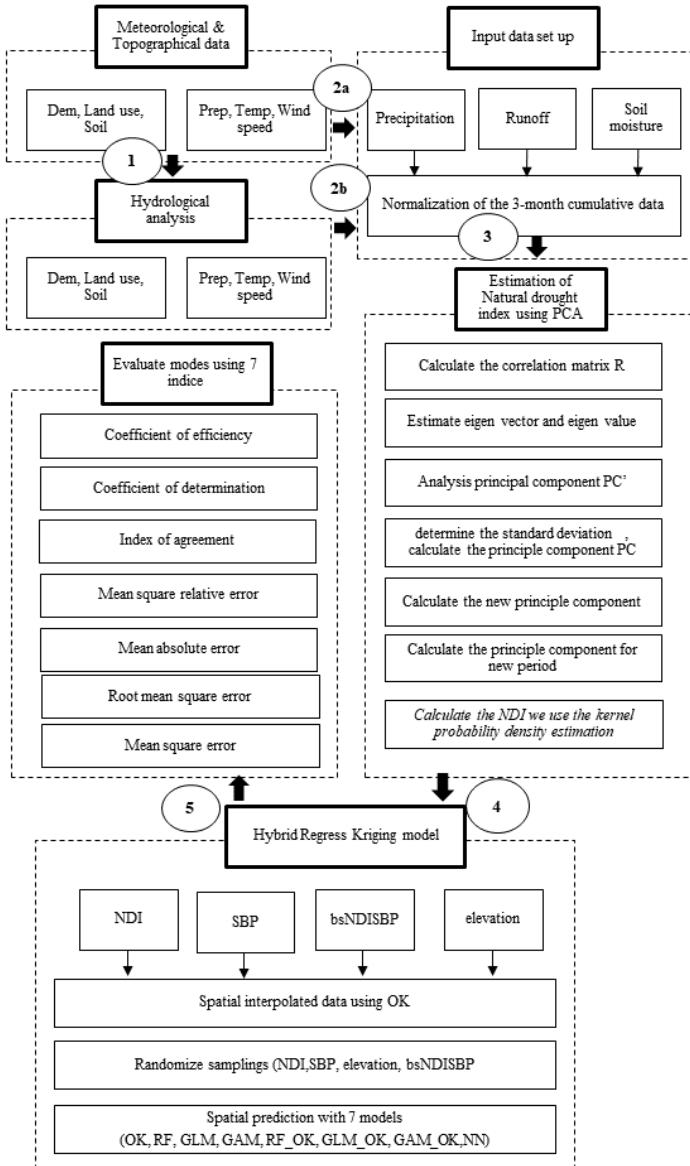


Figure 18. Hybrid egression kriging model to estimate spatial extreme drought

**Professor comments:**

- Review the objective and methods
- Clearly the method how to use ANN, SBP, NDI to improve spatial extreme drought coverage
- First concepts must be clear
- Professor does not want the initial results like that
- After finished concepts: how to use data, training data, estimate our results. It means that we need to clear each step before getting results
- Come back to read 2-3 weeks ago

## WEEK 15

### Clarify each steps of methodology

#### 1. Clarify methodology

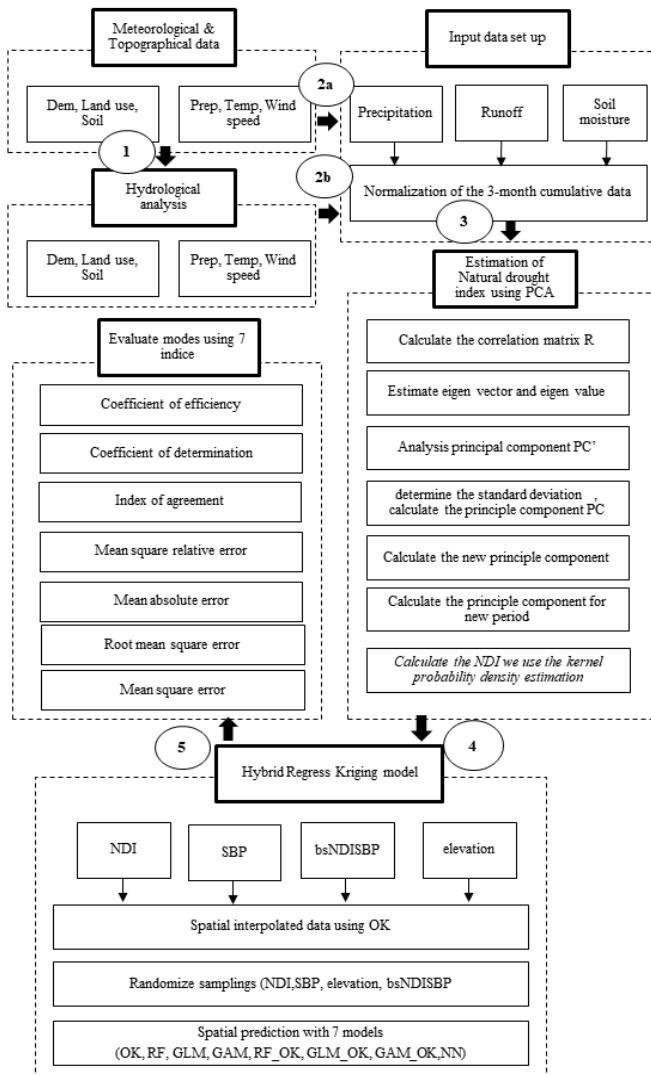


Fig 8. Framework estimate extreme drought coverage

Fig1 presents 5 main steps of framework estimate spatial extreme drought coverages. Step 1 and step 2 builds a hydrological model to extract runoff, soil moisture. Step 3 computes NDI. Step 4 setup the Neural Network Kriging, and final step 5 estimate results. The detail of processing was described bellow

**Commented [NTH4]:**

### Step 1: Prepare data for VIC model

The input data of VIC model are soil parameters, vegetation parameters, vegetation library, meteorological forcing. The outputs of VIC are soil moisture, evapotranspiration, runoff/streamflow, snow water equivalent. In our study, the digital elevation model (DEM) was provided by Mininstry of Land. The land cover map has spatial resolution at 100m interval. The soil map at scale 1:25000. The meteological data included 76 ASOS and 521 rainfall observation. Total 586 gridcell at 1/8<sup>0</sup> (12.5 km) was generated for VIC model. Maximum, minumum temperature and average wind speed was collected at each weather stations. They are interpolated to creat the grid with the same spatial 12.5 km (Son et al., 2011).

### Step 2: Calibration and validation VIC model

Ater prepared all input data, VIC model was calibrated and validated by comparing the stream flow at dam operation with observational discharge from Water Resource Maangement Infomation System (WMIS). The output of VIC model was extracted at each grid cell. They have a daily temporal scale. Then they were accumulated to monthly scale as these input data for computed NDI in the next step.

### Step 3: estimate natural drought index

NDI was computed by using principle component analysis (PCA). Precipitation, Runoff, Soilmoisture are the input of NDI computation. We computed NDI at 59 ASOS. That means ASOS provide the precipitation. The runoff and soil moisture are extracted from gridcells. These gridcells are matched to ASOS by closest neigborhood method. It notes that soil moisture and runoff are extraced for hourly scale, while the precipitation has the daily observation. They need to be accumulated to the montlly times-scale. PCA is reduce dimension methods. It based on the linearity that detemine the new vector that contains the most information from input vectors. The processing computation NDI follow by:

The data was standardized

$$x_{m,n} = \frac{o_{m,n} - \bar{o}_m}{\sigma_m} \quad (1)$$

The detemine the covariance matrix follow by:

$$R = \frac{1}{n-1} X \times X^T \quad (2)$$

The eigen value was determined follow by:

$$R.v = \lambda.v \quad (3)$$

The eigen vector was determine follow by:

$$PC' = v_1^T \times X \quad (4)$$

We change to the real vector by:

$$PC = \frac{PC'}{\sigma_{PC}} \quad (5)$$

We used eigen value above to computed for new periods

$$PC_{new}' = v_1^T \times X_{new} \quad (6)$$

$$PC_{new} = \frac{PC'_{new}}{\sigma_{PC'}} \quad (7)$$

The NDI was computed follow equation. It cumulates probability into stand variable follow approximate conversion developed in Abramowitz and Stegun (1965).

$$NDI = - \left( t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3} \right) \quad (8)$$

$$\text{If } 0.0 < P \leq 0.5 \text{ and } t = \sqrt{\ln(\frac{1}{P^2})}$$

Or the NDI was coputed follow equation:

$$NDI = \left( t - \frac{c_0 + c_1 t + c_2 t^2}{1 + d_1 t + d_2 t^2 + d_3 t^3} \right) \quad (9)$$

$$\text{If } 0.5 < P \leq 1.0 \text{ and } t = \sqrt{\ln(\frac{1}{(1-P)^2})}$$

Here,  $P$  is probability,  $c_0 = 2.515517$ ,  $c_1 = 0.802853$ ,  $c_2 = 0.010328$ ,  $d_1 = 1.432788$ ,  $d_2 = 1.89267$ ,  $d_3 = 0.001308$  is constain values.

This function is approximate standardized NDI

Depend the timescale of input precipitation, soil moisture, runoff intput data, we computed various NDI time-scale 1 month, 3 months, 6 months, 9 months, and 12 months. We chose 3 months timescale for drought analysis. The drought catalogue was classify similar SPI follow the table 1.

**Table 1.** Classification of NDI value

Values	Probability (%)	Drought category
--------	-----------------	------------------

$2.00 \leq \text{NDI}$	2.3	Extreme wet
1.99 ~ 1.50	4.4	Very wet
1.49 ~ 1.00	9.2	Moderately wet
0.99 ~ -0.99	68.2	Near normal
-1.00 ~ -1.49	9.2	Moderate drought
-1.50 ~ -1.99	4.4	Severe drought
$-2.00 \geq \text{NDI}$	2.3	Extreme drought

#### Step 4 : Artificial Neural network Kriging

Kriging is interpolation method that it weights the surrounding measured valued to derive a prediction for each location. However, the weights are based not only on the distance between the measured points and the prediction location but also on the overall spatial arrangement among the measured points. To use the arrangement in the weights, the spatial autocorrelation must be quantified. Normally Kriging model processing follow :(i) calculate the empirical semivariogram; (ii) fit model; (iii) create the matrices; (iv) make a prediction. For example,  $Z(s)$  is predicted NDI value at point at  $s$  ( $X, Y$ ):

$$Z(s) = \mu + \varepsilon(s) \quad (10)$$

Here,  $\mu$  is mean for the data (no trend), and error  $\varepsilon(s)$  with spatial dependence. Assume that the random error  $\varepsilon(s)$  is intrinsically stationary. The predicted NDI is formed as a weighted sum of the data

$$\hat{Z}(s_0) = \sum_{i=1}^N \lambda_i Z(s_i) \quad (11)$$

Here,  $\hat{Z}(s_0)$  is predicted NDI value at point  $s_0$ ,  $\lambda_i$  is unknown weight for the measured NDI value at the  $i$ th location,  $Z(s_i)$  is the NDI values at 59 ASOS location ( $N=59$ ). When making prediction for several locations, expect some of average, the difference between the predictions and the actual values should be zero. This is referred to as making the prediction unbiased. To ensure the predictor is unbiased for the unknown measurement, the sum of the weight  $\lambda_i$  must equal one. Using this constraint, make sure the difference between the true value,  $Z(s_0)$ , and predicted,  $\hat{Z}(s_0)$ , is as small as possible  $(Z(s_0) - \sum_{i=1}^N \lambda_i Z(s_i))^2$ . The solution to the minimization, constrained by unbiasedness, gives the kriging equation:

$$\Gamma \times \lambda = g \quad (12)$$

Here, gamma matrix  $\Gamma$  contains the modelled semivariogram values between all pairs of sample locations, where  $\lambda_{ij}$  denotes the modeled semivariogram values based on distance between the two samples identified as the  $i$ th and  $j$ th locations. The vector  $g$  contains the modelled semivariogram values between each measured location and the prediction location. The function is used to compute variogram follow equation:

$$\hat{\lambda}(h) = \frac{1}{2N(h)} \sum_{i,j} (Z(s_i) - Z(s_j))^2 \quad (13)$$

Here,  $\hat{\lambda}(h)$  is value of variogram at the lag  $h$  distance. The basic ideal of Kriging is finding the rule of changing the interested value follow the lag distance with measurements. The variogram characterizes the differences of data points depending on the distance between them.

To improve the spatial prediction, we use Regression Kriging (RK), one specific type of Kriging family. It considers the auxiliary to the interpolation processing. Regression kriging use a regression (Neural Network) to compute estimates for each predicted location, followed by ordinary kriging to interpolate the residuals. In our case, bias of gauge-based NDI and satellite-based precipitation was used as auxiliary variable of regression. The prediction of RK is again a weighted average:

$$\hat{Z}(s_0) = \sum_{k=0}^p \hat{\beta}_k \cdot q_k(s_0); q_0(s_0) = 1 \quad (14)$$

Here,  $q_k(s_0)$  are the values of the auxiliary variables at the target location,  $\hat{\beta}_k$  are the estimated regression coefficients and  $p$  is the number of predictor of auxiliary variables. RK combine these two approaches: regression is used to fit the explanatory variation and simple kriging with expect error value 0 to fit the residuals

$$\hat{Z}(s_0) = \hat{m}(s_0) + \hat{e}(s_0) = \sum_{k=0}^p \hat{\beta}_k \cdot q_k(s_0) + \sum_{i=1}^n \lambda_i \cdot e(s_i) \quad (15)$$

Here,  $\hat{m}(s_0)$  is the fitted drift,  $e(s_i)$  is the interpolated residual,  $\hat{\beta}_k$  are estimated drift model coefficients,  $\lambda_i$  are kriging weights determined by spatial dependence structure of residual, and  $e(s_i)$  is the residual at location  $(s_i)$ .

The residual  $e(s_i)$  is determined by regression using Neural network (NN) with data input (latitude, longitude, bias of NDI and SBP). Here, NN is used as function (Günther & Fritsch, 2010):

$$o(x) = f(w_0 + \sum_{j=1}^J w_j \cdot f(w_{0j} + \sum_{i=1}^n w_{ij} x_i)) = f(w_0 + \sum_{j=1}^J w_j \cdot f(w_{0j} + w_j^T x)) \quad (16)$$

Here,  $w_0$  is the intercept of the output neuron,  $w_{0j}$  is the intercept of the jth hidden neuron,  $w_j$  is the synaptic weight corresponding to the synapse starting at the jth hidden neuron leading to the output neuron,  $w_j = (w_{1j}, \dots, w_{nj})$  is the vector of all synaptic weights corresponding to the synapses leading the jth hidden neuron,  $x = (x_1, \dots, x_n)$ , and f is active function. We chose sigmoid active function because it is widely used in NN:

$$S(x) = \frac{1}{1 + e^{-x}} \quad (17)$$

**12.** In our study, 59 data points with attributes (longitude, latitude, bias NDI-SBP) are random subset to training (80%) and testing (20%). The number of hidden layer and number of neurons are optimizing by trial and minimize RMSE. Finally, the predicted NDI using equation (15) considered bias NDI-SBP as the adjustment.

#### Step 5: Evaluate models

To assess the effectiveness of the different modelling methods three indices were used: the mean absolute error (MAE), the root-mean-square error (RMSE) and relative root-mean-square error (RRMSE)

#### 2. Discussion

-Review all steps help us understanding our methodology. All information of **step 1** were attracted from papers. If I had original VIC model , I could be more understand this step. For example, I can know exactly the spatial distribution of 586 gridcells. Then the interpolated map would be better than from 59 stations. Addition, several useful data such as land cover, DEM with high spatial resolution could be re-used in present study. They are considered as the extra auxiliary for multivariate regression models. However, at this time, we should focus on methodology. Therefore, we will clarify our study based on 59 stations.

- **Step 4** of framework is the most sophisticated step. We proposal spend more time to understand this step. We would like to know which the easiest way is to combine kriging and NN. Beside that the data for training and testing of neural network is small. It is necessary to be enhanced. We believe that the NN give more accurate results if we have more reasonable data.

#### References

- Abramowitz, M., & Stegun, I. A. (1965). Handbook of mathematical functions with formulas, graphs, and mathematical tables. In *US Department of Commerce* (pp. 937): National Bureau of Standards Applied Mathematics series 55.  
 Günther, F., & Fritsch, S. (2010). neuralnet: Training of neural networks. *The R journal*, 2(1), 30-38.

Son, K.-H., Bae, D.-H., & Chung, J.-S. (2011). Drought Analysis and Assessment by Using Land Surface Model on South Korea. *Journal of Korea Water Resources Association*, 44(8), 667-681.  
doi:10.3741/JKWRA.2011.44.8.667

## WEEK 16

### Regression neural network kriging

The aim of this report to clarify regression neural network kriging (RNNKR) in step 4 and the evaluated models in step 5 of proposed framework (Fig 1)

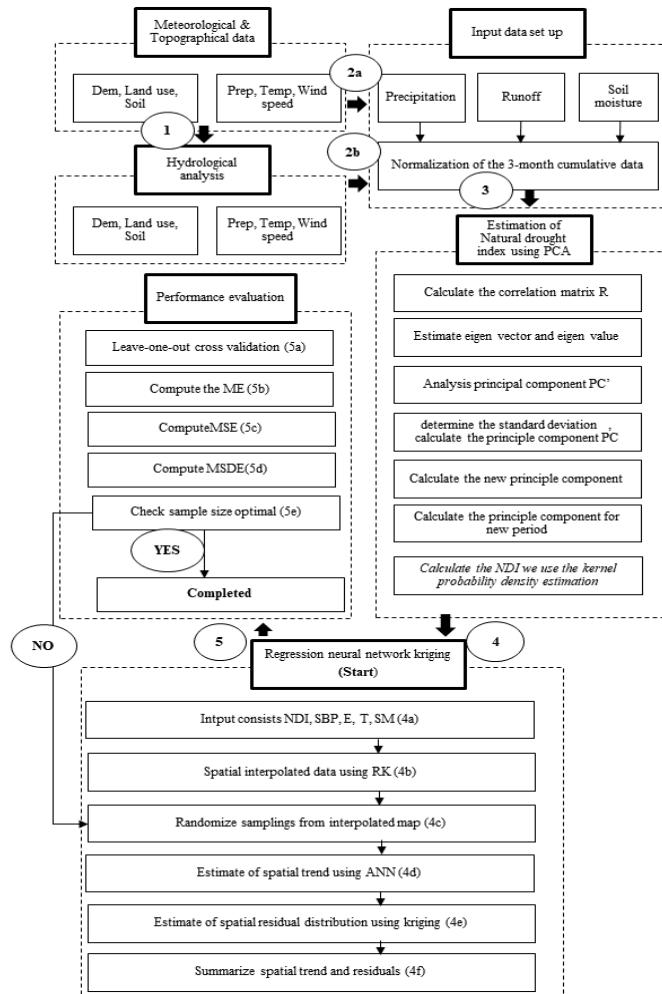


Fig 9. Framework estimate extreme drought coverage

## 1. Regression neural network kriging (step 4 of our framework)

Regression neural network kriging (RNNKR) is a spatial interpolated model combine regression kriging, artificial neural network (ANN) and residual kriging. The RNNKR model uses randomly sampled data from the interpolated map. Estimate spatial trend by using ANN. The RNNKR used both available data and random interpolated data as training dataset. The RNNKR models includes 4 main steps. Step 1 is spatial prediction using RK. Step 2 is spatial random sampling from interpolated map. Step 3 is estimate spatial trend using ANN. Step 4 is summarize spatial trend and residual in the interpolation. The whole process is repeated until the optimal sample size is found. The optimal sample size can be determined based on performance indices such as mean error (ME), mean square error (MSE), mean square standardized error (MSDE). The detail of each steps follows below:

**Step 1:** NDI at 59 ASOS locations is standardized and merge to auxiliary data from remote sensing: satellite-based precipitation (SBP), elevation (E), temperature (T), soil moisture (SM). The auxiliary data is converted from original raster's in Geo Tiff format to points. Overlap with NDI points by using spatial interpolation regress kriging. Regression matrix includes available NDIs as observational data, while auxiliary data (SBP, E, T, SM) are used as predictors. The RK is used to interpolate NDI values for whole study area.

**Step 2:** The sample data is random split to 10 subsets follow the ratio 80% for training and 20 % for testing.

**Step 3:** For estimate the spatial trend, ANN is used. A simple ANN structure with 1 input layer, 1 hidden layer and 1 output layer was utilized to determine the spatial trend (Fig 2).

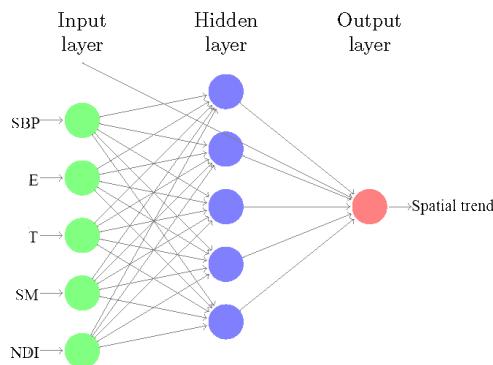


Fig 10. Structure of ANN

The input layer has 5 nodes: satellite-based precipitation (SBP), elevation (E), temperature (T), soil moisture (SM), NDI. Number of hidden nodes in the hidden layer was determined by using a trial and minimize error. The optimize number of notes is selected from 1 to 30 or until the ANN performance has no longer improvement. The standard back-propagation algorithm was used to train ANN. The logistic sigmoid activation function was chosen to compute output of each neuron.

$$S(x) = \frac{1}{1 + e^{-x}} \quad (1)$$

**Step 4:** Estimate of spatial residual distribution using variogram. The empirical variograms were computed for kriging models, and parameter, correlation, range, and sill variance. The operation least square method is used. The Gaussian model was selected to fit the empirical variogram model. The variogram model can be described follow Oliver and Webster (1990):

$$\gamma(h) = c \left\{ 1 - e^{-\frac{h^2}{a^2}} \right\}, h \geq 0 \quad (2)$$

Here,  $\gamma$  is the semivariance,  $h$  is lag distance,  $c$  is the sill variance, and  $a$  is the range of correlation.

**Step 5:** The leave one out cross validation is used to evaluate model's performance. MAE and RMSE is used in leave one out cross validation. Cross validation process consists of:

(i)Obtaining the kriging prediction  $\hat{Z}(s_i)$  at each sample point  $s_i$  i=1,...,n (as if the sample value at such points were unknown) from the observations at the n-1 remaining points ( or from a set of neighboring observations). The prediction variance at each sample point  $\hat{\sigma}^2(s_i)$  is also computed.

(ii)Calculating the following diagnostic statistics from the results obtained in (i):

The mean prediction error:

$$ME = \frac{1}{n} \sum_i^n (Z(s_i) - \hat{Z}(s_i)) \quad (3)$$

the mean squared prediction error:

$$MSE = \frac{1}{n} \sum_i^n (Z(s_i) - \hat{Z}(s_i))^2 \quad (4)$$

the mean squared standardized prediction error:

$$MSDE = \frac{1}{n} \sum_i^n \left( \frac{Z(s_i) - \hat{Z}(s_i)}{\hat{\sigma}(s_i)} \right)^2 \quad (5)$$

The semivariogram choice is successful when ME should be approximate 0, which is indicative of non-systematic prediction error, no matter what the fitted semivariogram is, kriging prediction are unbiased, and the ME is expected to tend to 0. MSE should be small, and MSDE should be approximately 1.

### 13. 2. Discussion

-Several studies (dos Santos, 2020; Hengl et al., 2007; Seo et al., 2015; Yuan et al., 2020) showed the regression kriging is outperformance compare to conventional interpolation methods (IDW, Kriging). Because it is sophisticated with considering covariances of systematic of observational and auxiliary data from remote sensing. Remote sensing data have better spatial coverage than gauge-based data. In addition, ANN is used to determine the spatial trend with non-linear approach is also support the accuracy improvement. However, some limitation of regression and ANN should be considered:

(1) *Data quality*: RK relies completely on the quality of data. If the data comes from different sources and have been sampled using biased or unrepresentative design, the predictions might be even worse than with simple mechanistic prediction techniques. Even a single bad data point can make any regression arbitrarily bad, which affects the RK prediction over the whole area.

(2) *Under-sampling*: For regression modelling, the multivariate feature space must be well-represented in all dimensions. For variogram modeling, an adequate number of point-pairs must be available at various spacings. Webster and Oliver (2007) recommend at least 50 and preferably 300 points for variogram estimation. Hengl et al. (2007) strongly recommend using RK only for data sets with more than 50 total observations and at least 10 observations per predictor to prevent over-fitting.

(3) *Extrapolation outside the sampled feature space*: If the points do not represent feature space or represent only the central part of it, this will often lead to poor estimation of the model and

poor spatial prediction. This is especially important for linear modelling where the prediction variance exponentially increases as we get closer to the edges of the feature space. For this reason, it is important that the points be well spread at the edges of the feature space and that they be symmetrically spread around the center of the feature space.

-Theory shows our framework could be used to estimate extreme drought coverage. However, results depend on data and setup the corrected methodology. We propose spending time to make

our methodology is clear and concise. For example, should we use other method to fit variogram, or use other active functions, or add more easting, northing value in predictors?

## References

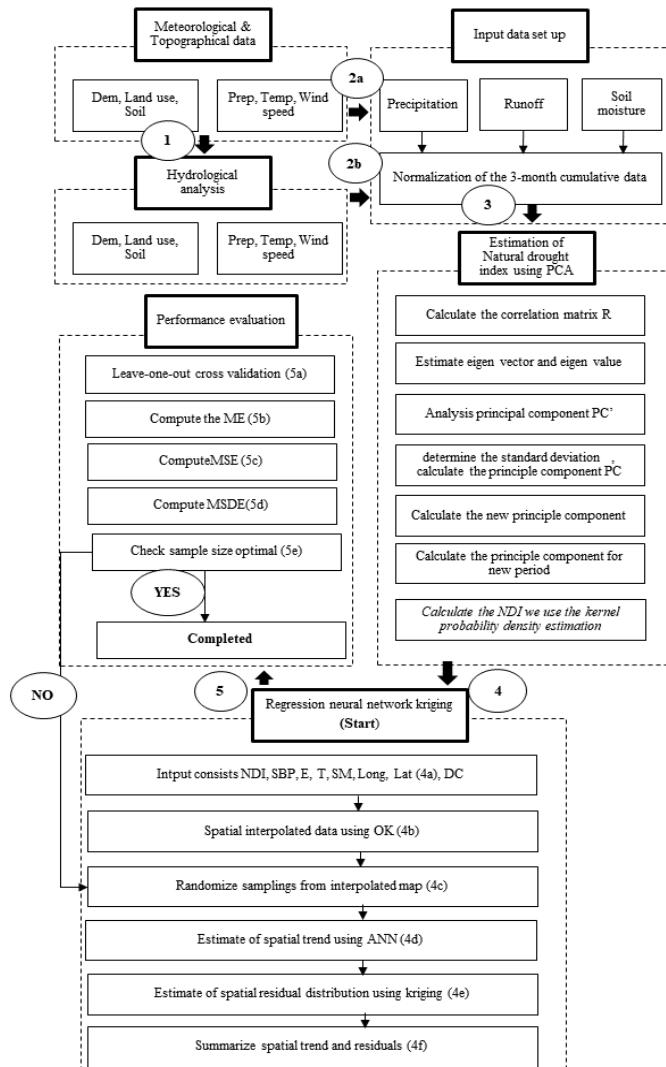
- Dos Santos, R. S. (2020). Estimating spatio-temporal air temperature in London (UK) using machine learning and earth observation satellite data. *International Journal of Applied Earth Observation and Geoinformation*, 88, 102066. doi:<https://doi.org/10.1016/j.jag.2020.102066>
- Hengl, T., Heuvelink, G. B. M., & Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. *Computers & geosciences*, 33(10), 1301-1315. doi:<https://doi.org/10.1016/j.cageo.2007.05.001>
- Oliver, M. A., & Webster, R. (1990). Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System*, 4(3), 313-332.
- Seo, Y., Kim, S., & Singh, V. P. (2015). Estimating spatial precipitation using regression kriging and artificial neural network residual kriging (RKNNRK) hybrid approach. *Water resources management*, 29(7), 2189-2204.
- Webster, R., & Oliver, M. A. (2007). *Geostatistics for environmental scientists*: John Wiley & Sons.
- Yuan, Q., Xu, H., Li, T., Shen, H., & Zhang, L. (2020). Estimating surface soil moisture from satellite observations using a generalized regression neural network trained on sparse ground-based measurements in the continental U.S. *Journal of Hydrology*, 580, 124351. doi:<https://doi.org/10.1016/j.jhydrol.2019.124351>

Professor comments:

-Clarify step 4 of framework

## WEEK 17

**A framework evaluates extreme drought coverage using natural drought index, Satellite base precipitation and artificial neural network**



*Fig 11. Framework estimate extreme drought coverage*

## 1. Clear and concise our methodology

We added the spatial trend analysis, correlation, and multi-collinearity test and describe the detail random sampling method.

### 1.10 Spatial trend analysis

Estimate spatial trend to understand the homogeneous spatial data. The spatial trend raises the points above a plot of the study site to the height of the values of the attribute of interest in a three-dimensional plot. The points are then projected in two directions (by default, north and west) onto planes that are perpendicular to the map plane. A polynomial curve fit to each projection. The entire map surface can be rotated in any direction, which also changes the direction represented by the projected planes. If the curve through the projected points is flat, no trend exists. Fig 2 shows NDI on August 2015 has the trend increase from the North-West, highest in Middle and decrease toward the South-East.

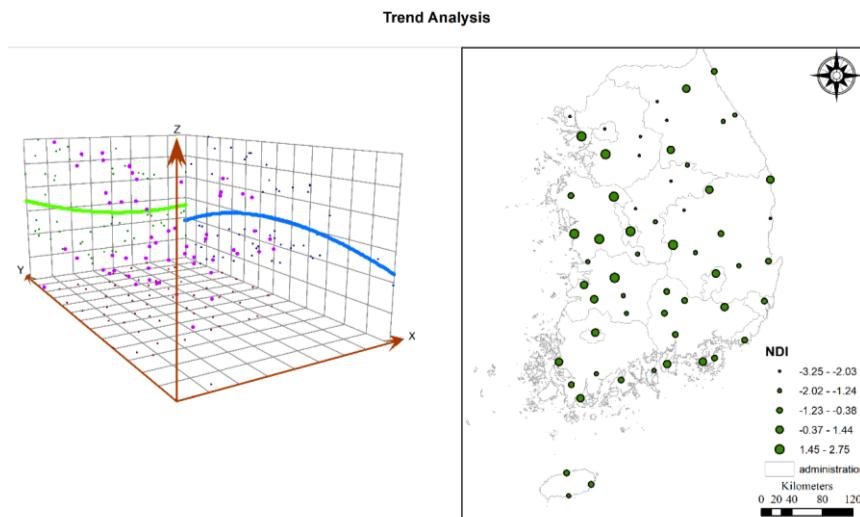


Fig 12. Spatial trend of NDI on August 2015

### 1.11 Multi-collinearity

Estimate correlation and the independent of variables helps to determine the predictors in the regression model. The predictors were chosen based on physics and statistics. SBP (satellite-based precipitation), SM (soil moisture), T (temperature), E (elevation), X (longitude), Y (latitude) could impact to drought. Other parameter is CD (the distance from location to coastal boundary) was

also considered. Because the coastal climate is various to inland climate. Multicollinearity was used to check the independence of variables (O'brien, 2007).

$$\text{tolerance} = 1 - R_j^2 \quad (1)$$

$$VIF = \frac{1}{\text{tolerance}} \quad (2)$$

Here,  $R_j^2$  is the coefficient of determination of independent variable  $j$  on all other independent variables. A tolerance of less than 0.2 or 0.10, or VIF of 5 or 10 and about shows a multicollinearity. Number of variables were chosen following the multi-collinearity condition.

### 1.12 Random sampling

The number of random sampling affects to result of ANN. Increase the size of sample support the training and testing is more accurate. But the time-consuming for computing is also increasing. Therefore, we should optimize the number of sampling by trial and loop. Data were random sampling from interpolated map. We create random points with spatial space from 10 km, 5 km and 2.5 km approximately equal 0.1, 0.05 and 0.025 degrees. The random points take values from interpolated raster by using overlapping with the closest pixel. We used mean error (ME) to choose the suitable random sampling size:

$$ME = \frac{1}{n} \sum_i^n (Z(s_i) - \hat{Z}(s_i)) \quad (3)$$

The suitable random sampling size that has the lowest value of ME. ME was computed based on comparison 59 NDIs and predicted values. Fig 3 presents the various random sampling from kriging interpolation maps.

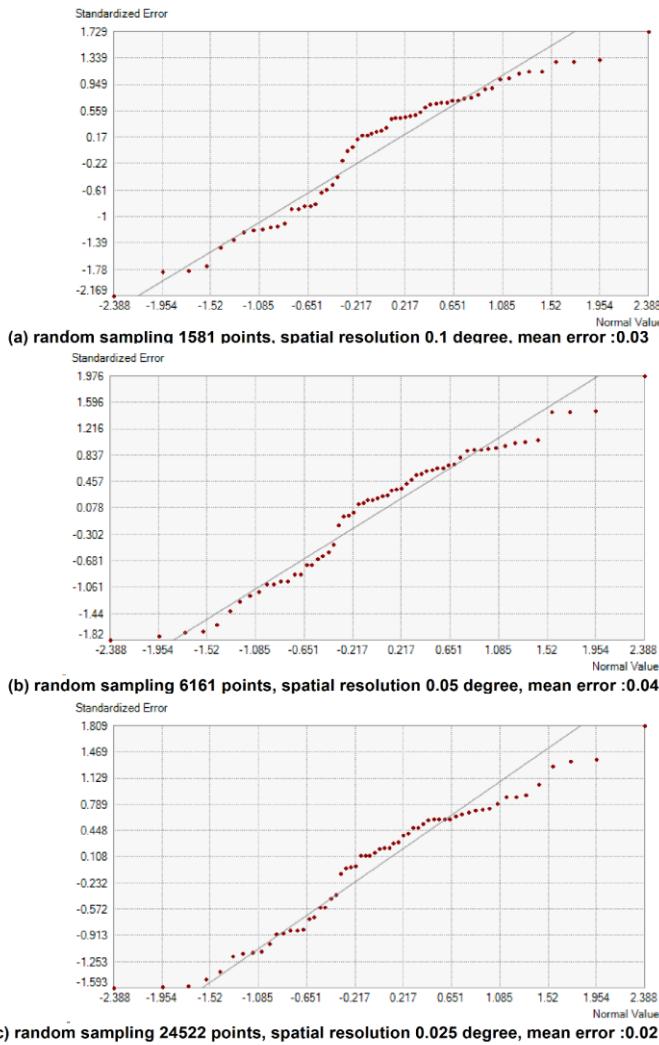


Fig 13. Various sampling number sizes

## 2. Discussion

(1) Analysis spatial trend of data determines the homogeneous of data. Based on it, we can avoid to over smooth surface problems which often occurs when data is interpolated. For example, our data has the trend high value in middle, and lower value to North-West and South-East. It should be considered when compare with final interpolation when it included the auxiliary data. It is one of necessary thing for discussion.

Two dis-transparent issues: the number of auxiliary data and the number of random sampling processing were clear. Using multi-collinearity can determine the independent of variables. Its help to reduce the number of auxiliary data. The number of random sampling can be chosen by mean error (ME) that compared 59 available NDIs and predicted values. However, the results show that the accuracy does not increase linearity with several random sampling sizes. Therefore, a number of sampling were chosen based on experiments. It depends on a specific data.

(2) One issue should be considered that spatial 59 NDI points do not fit in administrative boundary of South Korea. It leads to extrapolate values in the edge boundary. The consequences of this could make the spatial dependence change both the distance and the direction-anisotropy characteristic of data. Therefore, the uncertainty of the predicted spatial map should be added to make an outcome more valuable.

(3) Next step, we propose using available data to analyze results. After getting drafting results, we can adjust some parts of methodology if it is not suitable. In another direction, we should spend more time to understand fully our methodology before analysis. For instance, which is the best method to integrate residual of kriging and ANN? Beside using Gaussian function to fit variogram, why we do not use other functions?

## References

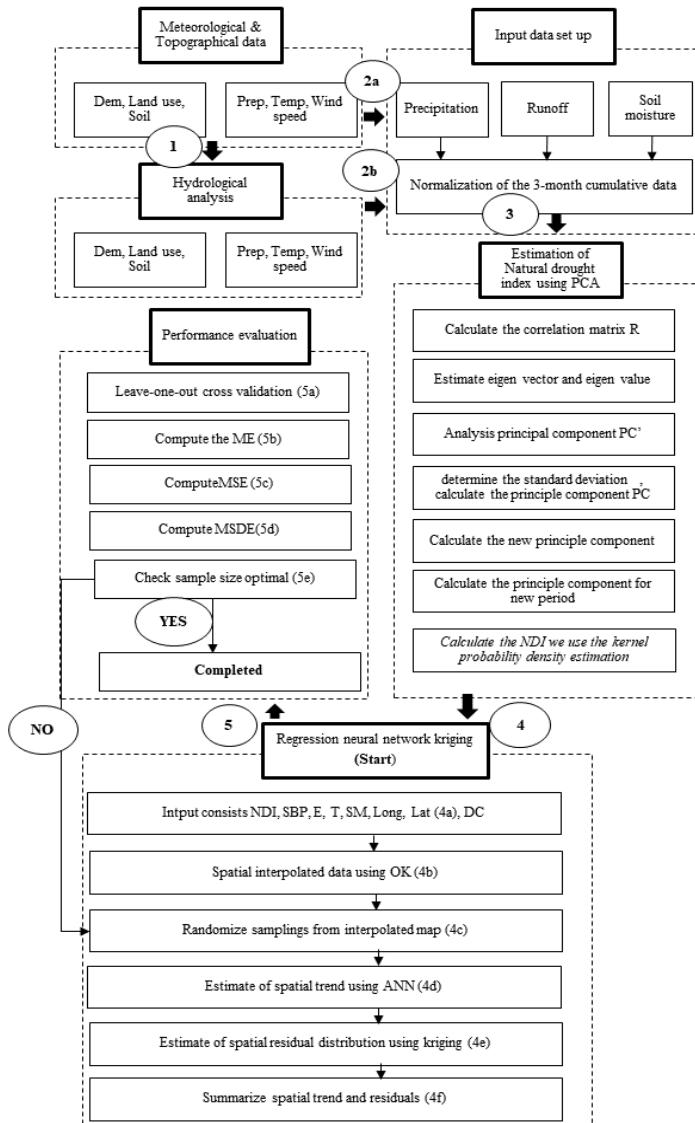
- O'brien, R. M. (2007). A caution regarding rules of thumb for variance inflation factors. *Quality & quantity*, 41(5), 673-690. 1

**Professor comments:**

I have not time to read your report, just follow the schedule.

## WEEK 18

**A framework evaluates extreme drought coverage using natural drought index, Satellite base precipitation and artificial neural network**



*Fig 14. Framework estimate extreme drought coverage*

## 1. Combine Neural network and Kriging

The aim of this report is reviewing the methodology of integrated the Artificial Neural Network (ANN) and interpolation Kriging (KR). We compare previous studies to check the reasonableness of our method. The framework estimates spatial precipitation using regression kriging and artificial neural network residual kriging (RKNNRK) is the most match our objective (Seo et al., 2015). The flowchart of RKNNRK is presented in Fig 2.

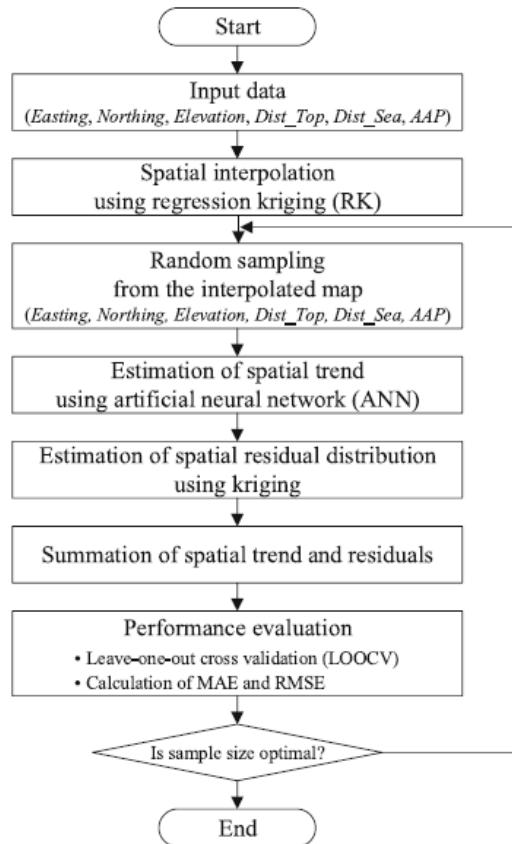
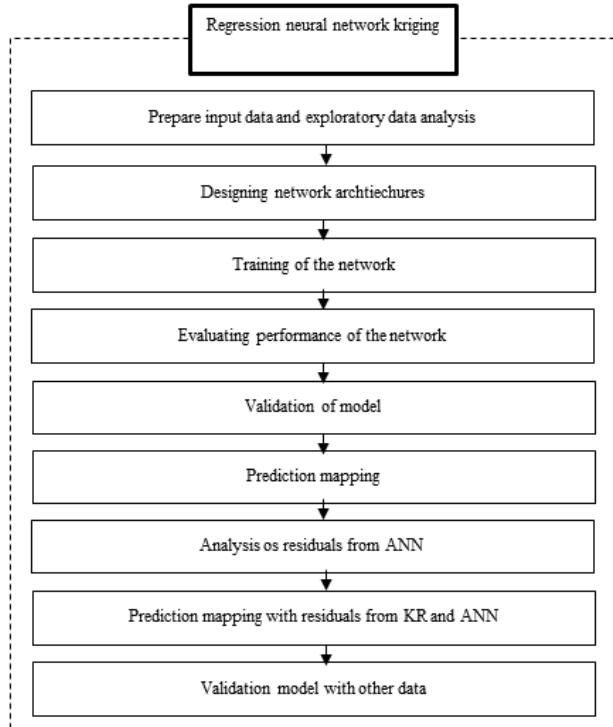


Fig 15. Flowchart of RKNNRK

The main idea of this study is using various auxiliary to improve spatial distribution of average annual precipitation for JeJu island, South Korea. ANN was used to estimate spatial trends. This idea similar to the study analyzes surface contamination by radionuclides after nuclear accident on Chernobyl Nuclear Power Plan (Kanevsky et al., 1996). Backpropagation training algorithm which is a supervised learning algorithm was applied. The residuals were analyzed and in case of

correlated residual Geostatistics was used. It means that ANN are used for training to predict the surface contamination and compare to the observational value to estimate the difference (residuals). Methodology processing includes 9 steps present in Fig 3.



*Fig 16. Framework for predicting spatial radionuclides from Chernobyl nuclear power plant.*

The problem that there are still uncertainties about accident scenario and details on physical and chemical composition of the time dependent source term, wind and rain fields at different scales, etc. Moreover, atmospheric dispersion models nonlinearly depend on many parameters (wet and dry deposition velocities, boundary layer parameterizations, orography, etc.). Measurements are used to estimate them. It is not evident that the use of atmospheric dispersion models should lead to the stationary residuals. This work is based on a simple idea. If data represent large scale trends over the entire region and small-scale variability, try to estimate nonlinear trend with ANN and then analysis residuals.

Results shows that NNRK model can be used in case of complex nonlinear trends over the entire region and small-scale spatial variability.

In the other idea is using Kriging-based model to improve rainfall accuracy (Bae, 2013) was presented in Fig 4. We name this study as study 4.

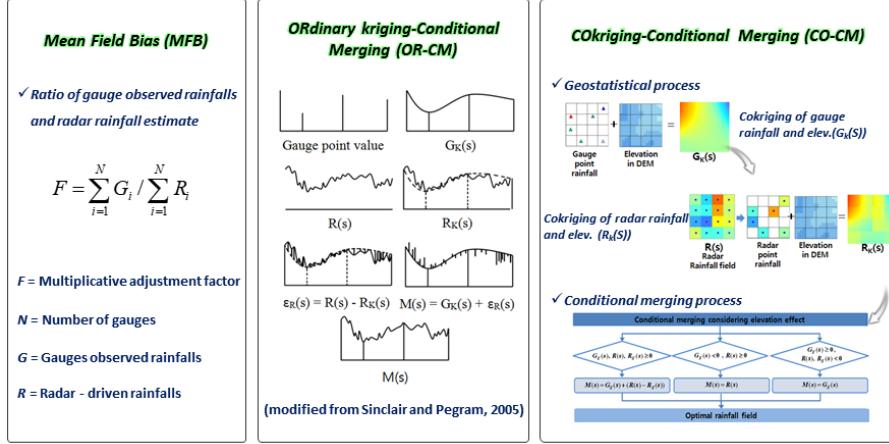


Fig 17. Adjustment methods for radar rainfall estimate

The basic concepts from this study is using Kriging to extract the optimal information content from observational gauges rainfall. A mean field based on the Kriged rain gauge data is adopted, while the spatial detail from radar is retained, reducing bias, but keeping the spatial variability observed by the radar. The variance of the estimate is reduced in the vicinity of the gauges where they are able to provide good information on the true rainfall field (Sinclair & Pegram, 2005).

Overview of technique conditional merging process is presented in Fig 5.

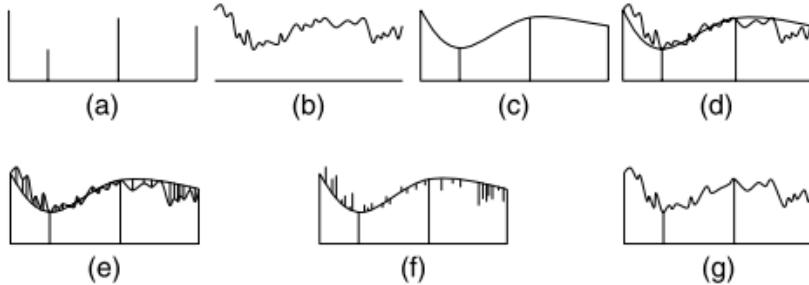


Fig 18. The conditional merging processes. (a) The rainfall field is observed at discrete points by rain gauges.

(b) The rainfall field is also observed by radar on a regular, volume-integrated grid.

(c) Kriging of the rain gauge observations is used to obtain the best linear unbiased estimate of

*rainfall on the radar grid. (d) The radar pixel values at the rain gauge locations are interpolated onto the radar grid using Kriging. (e) At each grid point, the deviation C between the observed and interpolated radar value is computed. (f) The field of deviations obtained from (e) is applied to the interpolated rainfall field obtained from Kriging the rain gauge observations. (g) A rainfall field that follows the mean field of the rain gauge interpolation, while preserving the mean field deviations and the spatial structure of the radar field is obtained.*

This technique uses radar observations of rainfall fields to estimate the errors associated with using Kriging to interpolate the rain gauge observations and condition the Kruged gauge field. The spatial detail of the final merged field is improved while maintaining the mean field characteristics measured by the gauges.

## **2. Discussion**

- (1) The different between study spatial precipitation for JeJu island (study 1) and previous study about contamination environment from Chernobyl (study 2) in the input data and sampling data. Study 1 had extended variables predictors (easting, northing, distant to top elevation, distance to coastal). Meanwhile study 1 used only easting (X), northing data (Y). Study 2 use random sampling to create various number of input data for training. This step was ignored in study 1.
- (2) We found that we can using ANN as study to gain the hidden information from data similar to study 1, using multivariable auxiliary to support training and testing data as in study 2. Study 4 also like our study. Instead of rainfall, we estimate NDI value for South Korea. Radar rainfall was replaced by satellite-based precipitation (SBP). The key finding from study 4 that using conditional merging of gauge-based data and satellite-based data. We propose using this technique for conditional merging residuals from ANN model and residual from kriging.
- (3) In the short, our study based on gauge-based NDI, satellite-based precipitation, Kriging model, ANN, conditional merging residuals.

(4) In the next step of our study, we propose make a plant for processing data, write codes, analyze results.

## References

- Bae, D. H. (2013). Adjustment methods for radar rainfall estimate. Weather Obs. & its Application Workshop on WISE Platform.
- Kanevsky, M., Arutyunyan, R., Bolshov, L., Demyanov, V., & Maignan, M. (1996). Artificial neural networks and spatial estimation of Chernobyl fallout. Geoinformatics, 7(1-2), 5-11. doi:10.6010/geoinformatics1990.7.1-2\_5
- Seo, Y., Kim, S., & Singh, V. P. (2015). Estimating spatial precipitation using regression kriging and artificial neural network residual kriging (RKNNRK) hybrid approach. Water resources management, 29(7), 2189-2204.
- Sinclair, S., & Pegram, G. (2005). Combining radar and rain gauge rainfall estimates using conditional merging. Atmospheric Science Letters, 6(1), 19-22. doi:10.1002/asl.85

## WEEK 19

### **A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network**

#### **1. Expected results**

We need to determine expected results before doing analysis. Some figures will be used follow the list:

- Fig 1. Methodology
- Fig 2. Study area, location of ASOS
- Fig 3. Time series of the average mean area of NDI from 1981 to 2016. It is used to determine the specific time when occurred the highest average mean area NDI. This specific time are used to analyze spatial drought coverages
- Fig 4. 3D chart spatial trend of NDI at the specific time
- Fig 5. Spatial distribution of NDI, satellite-based precipitation, soil moisture, temperature, elevation, Longitude, Latitude
- Fig 6. Covariance of NDI, satellite-based precipitation, soil moisture, temperature, elevation, Longitude, Latitude
- Fig 7. ANN structure of training algorithm.
- Fig 8. Cokriging-conditional merging variogram of NN and Kriging
- Fig 9. The predicted Map and uncertainty map

We will provide 1 table:

-Table 1. Evaluate model using cross-validation

## 2. Time schedule

We are going do analysis with 3 weeks. The detail of plan is present in table 1 below

	<b>Contents</b>	<b>Sub-contents</b>
Week 1	□ Data processing	<ul style="list-style-type: none"> <li>①. Convert NDI to spatial-space format</li> <li>②. Compute monthly average</li> <li>③. Determine the specific time of the maximum value NDI</li> <li>④. Retrieved precipitation from satellite</li> <li>⑤. Retrieved temperature from satellite</li> <li>⑥. Retrieved soil moisture from satellite</li> <li>⑦. Retrieved elevation from satellite</li> <li>⑧. Covert satellite-based data to grid-point and grid cell format</li> <li>⑨. Analysis spatial trend of NDI</li> </ul>
Week 2	□ ANN model	<ul style="list-style-type: none"> <li>⑩. Create surfaces for NDI, satellite-based data, longitude, latitude follow 3 spatial resolutions: 0.1;0.05;0.025 degree</li> <li>⑪. Extract satellite-based precipitation, soil moisture, temperature, elevation, longitude, latitude as 59 ASOS location</li> <li>⑫. Set up the neural network regression (NNR) to predict NDI from auxiliary variables above</li> <li>⑬. Build the structure of ANN</li> <li>⑭. Training and validate ANN</li> <li>⑮. Optimize structure of ANN</li> <li>⑯. Create table to evaluate model</li> </ul>

	<b>Contents</b>	<b>Sub-contents</b>
Week 3	□ Merging-condition merging model	<p>⑯. Compute the residual 1 of ANN</p> <p>⑰. Compute residual 2 of Kriging regression</p> <p>⑱. Compute Cokriging-conditional merging variogram of NN and Kriging</p> <p>⑲. Analysis spatial NDI coverage by using conditional variogram</p>

### 3. Discussion

- (1) It is important to set up the map's projection for our study. Because the various projections could lead to the bias when we overlap multiple map layers. The satellite-based data follow the world global system (WGS). These layers should be projected for the specific coordinate system. Unfortunately, we have not known about official coordinate system of South Korea. It should be considered before doing analysis.
- (2) The monthly satellite-based data takes a lot memory of drive. Because our study covers whole South Korea with the highest spatial resolution at 0.025 degree ~ 2.5 km. The capacity of hard drive should be extended for better analyze data performance. Especially, it impacts to training model of ANN if we have an insufficient memory.

**Professor comments:**

- Determine the OS using for this analysis. We can use the linux service if it is necessary. In case using personal computer, we can buy a new hard drive.
- It is technique issue. How to fix the coordinator system for South Korea.

## WEEK 20

A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network

### 1. Coordinate system of Korea

#### 3.1. What is the coordinate system?

A geographic coordinate system (GCS) uses a three-dimensional spherical surface to define locations on the earth. A GCS includes an angular unit of measure, a prime meridian, and a datum (based on a spheroid). Projected coordinate systems are any coordinate system designed for a flat surface such as a printed map or a computer screen.

The coordinate system is generally divided into two types: coordinate geographic system (GCS) and Projected Coordinate System (PCS). A geo-coordinate system, a coordinate system that uses a three-dimensional spherical to define a location on earth. A point in the geography coordinates system is specified as longitude and latitude height values. Longitude and latitude are measured angles from the center of the earth to the point on the earth's surface, often represented by degrees.

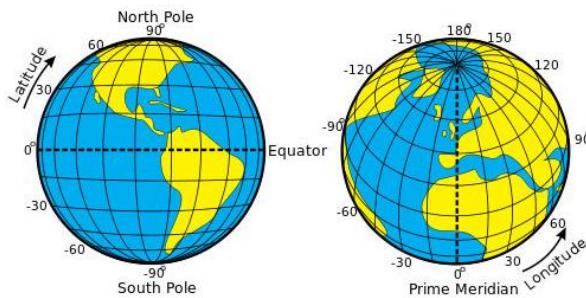


Fig 19. Geographic coordinate system (GCS)

Fig 1 shows the GCS of the earth. The line that is the basis of the latitude is called the Equator. It divides the earth into North and South.

Projected coordinate system that convert from a three-dimensional earth ellipse into a two-dimensional plane coordination system. It is called projection. In Fig 2 presents the Mercator projection concept.

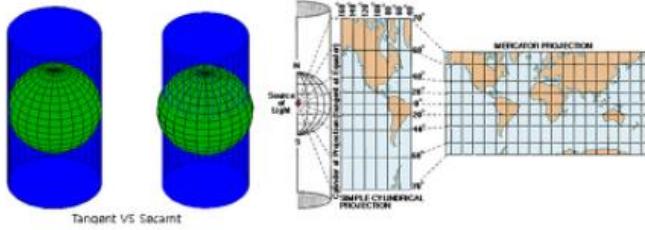


Fig 20. Mercator projection

TM coordinate system, UTM coordinate system, MGRS coordinate system, etc. are all called plane right angle coordinate system. Detail of coordinate system could be found in Van Sickle (2017).

### 3.2. Coordinate systems of South Korea

The popular GCS and PCS, which have been used in South Korea, are presented in table 1

Table 2. The popular GCS and PCS are used in South Korea

구분	단위체 종류	투영체계	원점		가산수치(m)		Units	Scale Factor	적용
			경도	위도	Eastings	Northings			
Geographic Coordinate Systems	Bessel 1841	-	0	0	-	-	Degree	-	해수부 등에서 일부 사용
Geographic Coordinate Systems	ITRF 2000(GRS80)	-	0	0	-	-	Degree	-	
Geographic Coordinate Systems	WGS 1984	-	0	0	-	-	Degree	-	Google, GPS
Projected Coordinate Systems	Bessel 1841	TM(동부원점)	129.0028902778	38	200,000	500,000	Meter	1	LMIS, KLIS, WAMIS 등
Projected Coordinate Systems	Bessel 1841	TM(중부원점)	127.0028902778	38	200,000	500,000	Meter	1	LMIS, KLIS, WAMIS 등
Projected Coordinate Systems	Bessel 1841	TM(서부원점)	125.0028902778	38	200,000	500,000	Meter	1	LMIS, KLIS, WAMIS 등
Projected Coordinate Systems	Bessel 1841	TM(KATEC)	128.0000000000	38	400,000	600,000	Meter	0.9999	네비업체
Projected Coordinate Systems	Bessel 1841	UTM(K)	127.5028900000	38	1,000,000	2,000,000	Meter	0.9996	세주소사업(행안부 자료)
Projected Coordinate Systems	Bessel 1841	UTM(51N)	123.0000000000	0	500,000	-	Meter	0.9996	
Projected Coordinate Systems	Bessel 1841	UTM(52N)	129.0000000000	0	500,000	-	Meter	0.9996	
Projected Coordinate Systems	ITRF 2000(GRS80)	TM(동부원점)	129.0000000000	38	200,000	600,000	Meter	1	국립지리원(현재 국가기준)
Projected Coordinate Systems	ITRF 2000(GRS80)	TM(중부원점)	127.0000000000	38	200,000	600,000	Meter	1	국립지리원(현재 국가기준)
Projected Coordinate Systems	ITRF 2000(GRS80)	TM(서부원점)	125.0000000000	38	200,000	600,000	Meter	1	국립지리원(현재 국가기준)
Projected Coordinate Systems	ITRF 2000(GRS80)	TM(KATEC)	128.0000000000	38	400,000	600,000	Meter	0.9999	서울 성북구청
Projected Coordinate Systems	WGS 1984	UTM(51N)	123.0000000000	0	500,000	-	Meter	0.9996	미국(군사용)
Projected Coordinate Systems	WGS 1984	UTM(52N)	129.0000000000	0	500,000	-	Meter	0.9996	미국(군사용)

Current Standard of coordinates of South Korea were proposed by The National Geospatial Intelligence Agency. The conversion from coordinate the Bessel 1841 ellipse (GCS) into the GRS80 follows the notice (2003-497). The coordinate systems impact the accuracy of the map. But the large-scale study area, the chosen projection coordinates system is less impact. Therefore, we can ignore this issue. In this study, we propose using the Korea\_2000\_Korea\_Central\_Belt as the map coordinate. All of map data will be assigned to this coordinate.

## 2. Determine Operation System

Window OS, Linux OS and web-based computation has diverged advantages. Window is the most polarity OS. Linux is the server system with high performance and high powerful computation. Cloud-computation is like the concept of using the server. The problem of it does not support fully graphical user interface. It takes the time to upload and download data. Editing on the server is inconvenient. Therefore, we propose using a personal computer (PC) for analysis. On the PC, we can create the virtual machine that is running Linux OS. Therefore, we can use both Windows and Linux at the same time. Because both Linux and Windows OS run on the same PC, it could be a little slow performance. In the recent, Window 10 has included the subsystem Linux directly on his OS. Detail of installation of subsystem Linux and python on Window 10 could be found in Davis (2017). Therefore, the problems of shared Random-Access Memory and Computer processors in the virtual machine can be handled. In addition, we can use server after completed preparation in personal computer if it is necessary.

## 3. Data processing

### 3.1. Multiple dimension data

We interpolated monthly NDI of 432 months (1981-2016) to create multiple raster files .The concept of a single and a multiple bands image is presented in Fig3.



Fig 21. The single band raster (left) and multiple bands raster (right)

We chose the network common data form (NetCDF) for this purpose (Rew & Davis, 1990). NetCDF is a file format for storing multidimensional scientific data such as temperature, humidity, wind speed and direction. Each of these variables can be displayed through a dimension (for instance time) by making a layer or table view from a netCDF file.

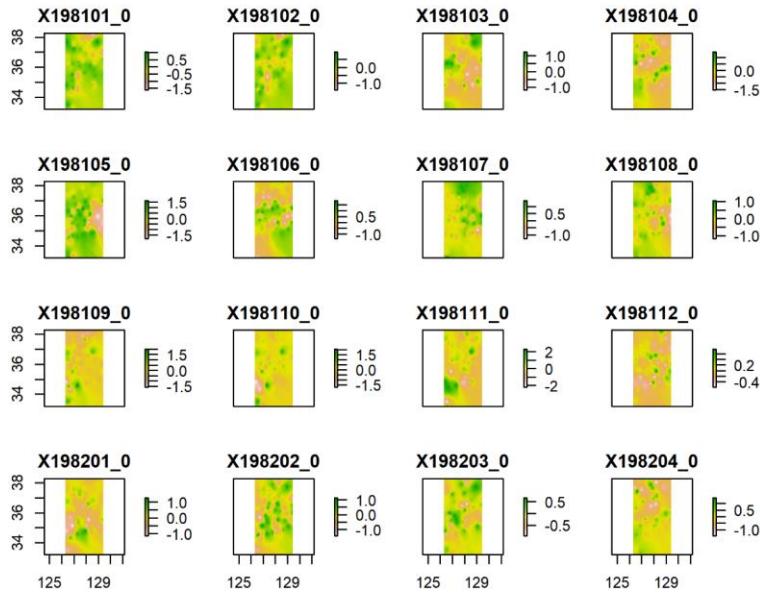


Fig 22. Analyze spatial drought using NDI from January 1981 to April 1982

### 3.2. Determine a specific time for analysis extreme drought coverage

There are several criteria to determine. We chose the time when the spatial distribution is highest fluctuation. Compare standard deviations of these images we determine September 2015 is the time that spatial drought has the most various coverages. The statistical analysis shows this period has a standard deviation of NDI at 1.89 and a mean of NDI at 0.07.

## 4. Discussion

Table 3. The status of processing data analysis

	Contents	Sub-contents
Week 1	□ Data processing	①. Convert NDI to spatial-space format ②. Compute monthly average ③. Determine the specific time of the maximum value <b>NDI</b> ④. Retrieved precipitation from satellite ⑤. Retrieved temperature from satellite ⑥. Retrieved soil moisture from satellite ⑦. Retrieved elevation from satellite

	Contents	Sub-contents
		<p>⑧. Covert satellite-based data to grid-point and grid cell format</p> <p>⑨. Analysis spatial trend of NDI</p>
Week 2	□ ANN model	<p>⑩. Create surfaces for NDI, satellite-based data, longitude, latitude follow 3 spatial resolutions: 0.1;0.05;0.025 degree</p> <p>⑪. Extract satellite-based precipitation, soil moisture, temperature, elevation, longitude, latitude as 59 ASOS location</p> <p>⑫. Set up the neural network regression (NNR) to predict NDI from auxiliary variables above</p> <p>⑬. Build the structure of ANN</p> <p>⑭. Training and validate ANN</p> <p>⑮. Optimize the structure of ANN</p> <p>⑯. Create a table to evaluate the model</p>
Week 3	□ Merging-condition merging model	<p>⑰. Compute the residual 1 of ANN</p> <p>⑱. Compute residual 2 of Kriging regression</p> <p>⑲. Compute Coking-conditional merging variogram of NN and Kriging</p> <p>⑳. Analysis spatial NDI coverage by using conditional variogram</p>

(1) To handle this study, we use several program languages (FORTRAN, MATLAB, Python, R) and both window OS, Linux OS. Addition, some unforeseen issues happens to lead the timetable changing. For instance, the determination of coordinate systems, finding the suitable operational system, specify the multiple band raster file format are time-consuming tasks. However, we can learn a lot of useful things from them. They are helpful for our future studies. Therefore, the study plan should be revised and extended.

(2) In the next week, we are going to complete the data processing section. It includes from task 4 to task 9. The main works will be successfully retrieved satellite-based data. How to select the suitable satellite-based data and what is the most effective way to retrieve them.

### **References**

- Davis, M. (2017). How to set up the Linux subsystem and install Twarc on Windows 10.
- Rew, R., & Davis, G. (1990). NetCDF: an interface for scientific data access. *IEEE computer graphics and applications*, 10(4), 76-82.
- Van Sickle, J. (2017). *Basic GIS coordinates* (Third edition ed.). Boca Raton: Taylor & Francis.

## WEEK 21

**A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network**

### 4. Data processing (continue)

*Table 4. The status of processing data analysis*

	Contents	Sub-contents
Week 1-2	<input checked="" type="checkbox"/> Data processing	<ul style="list-style-type: none"> <li>①. Convert NDI to spatial-space format</li> <li>②. Compute monthly average</li> <li>③. Determine the specific time of the maximum value NDI</li> <li>④. Retrieved precipitation from satellite</li> <li>⑤. Retrieved temperature from satellite</li> <li>⑥. Retrieved soil moisture from satellite</li> <li>⑦. Retrieved elevation from satellite</li> <li>⑧. Covert satellite-based data to grid-point and grid cell format</li> <li>⑨. Analysis spatial trend of NDI</li> </ul>
Week 3	<input type="checkbox"/> ANN model	<ul style="list-style-type: none"> <li>⑩. Create surfaces for NDI, satellite-based data, longitude, latitude follow 3 spatial resolutions: 0.1;0.05;0.025 degree</li> <li>⑪. Extract satellite-based precipitation, soil moisture, temperature, elevation, longitude, latitude as 59 ASOS location</li> <li>⑫. Set up the neural network regression (NNR) to predict NDI from auxiliary variables above</li> <li>⑬. Build the structure of ANN</li> <li>⑭. Training and validate ANN</li> <li>⑮. Optimize the structure of ANN</li> <li>⑯. Create a table to evaluate the model</li> </ul>
Week 4	<input type="checkbox"/> Merging-condition merging model	<ul style="list-style-type: none"> <li>⑰. Compute the residual 1 of ANN</li> <li>⑱. Compute residual 2 of Kriging regression</li> <li>⑲. Compute Cokriging-conditional merging variogram of NN and Kriging</li> <li>⑳. Analysis spatial NDI coverage by using conditional variogram</li> </ul>

### *3.3. Retrieved precipitation*

The precipitation was retrieved from IMERG. It has a spatial resolution at 0.1 degrees and tempo-resolution at monthly-scale. Fig 1 presents the spatial precipitation coverage using average monthly time scale. The lowest precipitation occurred in Chungcheong province with precipitation volume from 28 to 42 mm.

### *3.4. Retrieved temperature*

The temperature was retrieved from FLDAS Noah Land Surface Model. This model used data from meteorological satellite MERRA-2 and CHIRPS. It has the same tempo-spatial resolution with precipitation from IMERG. Chungcheong province has the highest temperature from 294 to 295 Kevin temperature approximates 22 to 23<sup>0</sup> C (Fig 2). It supports that this area has less precipitation and higher temperature than others.

### *3.5. Retrieved soil moisture*

The anomaly of Soil moisture content 0-10 cm underground) was also collected from FLDAS Noah Land Surface Model. Chungcheong province has the soil moisture at lower level at -0.17 to -0.08. The negative value shows the less soil moisture compare to average.

### *3.6. Retrieved elevation*

The elevation was retrieved from satellite STRM. Fig 3 shows that South Korea has a mountainous topography. It has elevation is almost higher than sea water level more than 10 meters. It does show clear different topographical characteristic of Chungcheong province compare to others. Spatial resolution of STRM is higher at 30m. Therefore, we need to convert to grid points has the same spatial resolution in the next step.

### *3.7. Convert satellite-based data to grid point and grid cell format*

The study area is divided into 3338 points as the centroid of each grid cell (Fig 5). Grid cells have spatial resolution 0.1 degree. These points included the South Korea and surroundings.

The satellite-based data were collected and transformed to grid cells (Fig6). The location and the spatial resolutions of them are the same. Be aware of the extension of interpolation to cover from efficiency data could lead to the uncertainty. The grid cells outside of administrative boundary has the most bias value. We should remove them to analyze spatial trend accurately.

### *3.8. Analysis spatial trend of NDI*

The Trend uses a global polynomial interpolation for fitting a smooth surface. The pseudo surface is defined by a mathematical function (a polynomial) to the input sample points. The trend surface changes gradually and captures coarse-scale patterns in the data. Fig 6 shows the trend of NDI that fitted with polynomials. Gyeoggi province has high NDI, and NDI value is low at

Chungcheongbuk province. Then it has increase forward to Daegu city. The outside of administration has strong bias lowest value. It is the cause of over extrapolated data.

## 5. Discussion

- (1) The extrapolated data could lead to strong bias results. In our study 59 ASOS location does not cover fully of South Korea. When it is interpolated, some edge of the image is not fulfilled. For instance, the Easter and the Wester of interpolated image has not available data (Fig 7). To resolve this issue, we need to add more data or determine the boundary of the study again. The other method that fits the image by extension (Fig 6). Although decrease the inaccuracy, it is the simplest method to handle the extrapolation.
- (2) Diverse spatial resolutions can impact to the results. The finer spatial resolution leads to the more accurate results. However, it is depending on the available data. In our study with 59 ASOS has interval space, approximate 12 km, spatial resolution of 0.1 degrees is suitable. Therefore, we propose to use only 0.1 degrees spatial resolution in the next analysis.

## Appendix

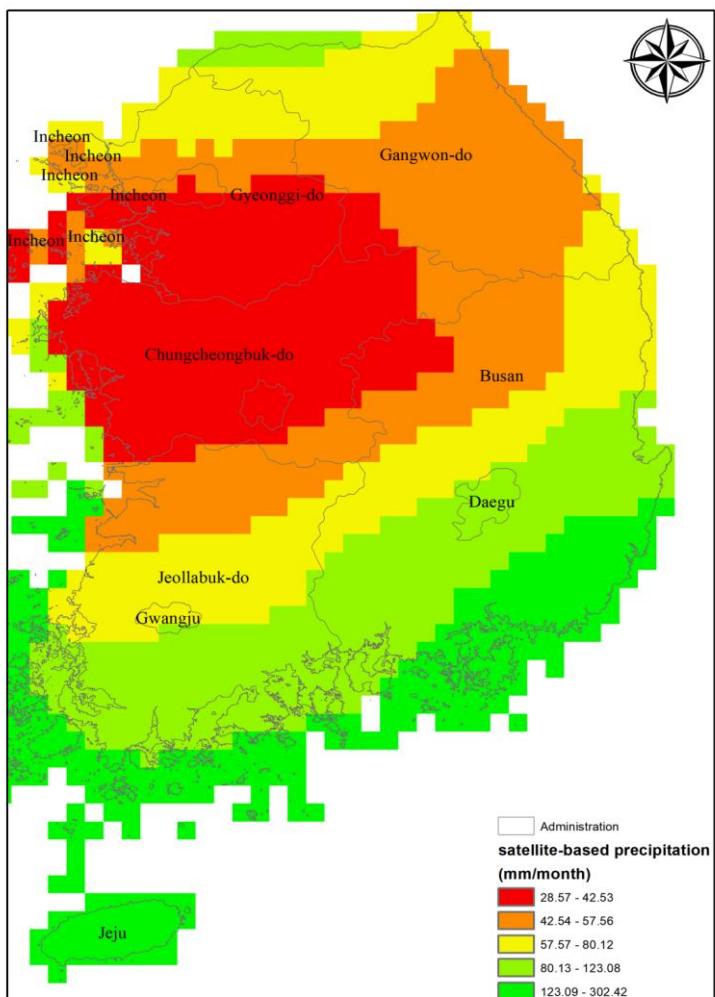


Fig 23. Spatial precipitation coverages on September 2015

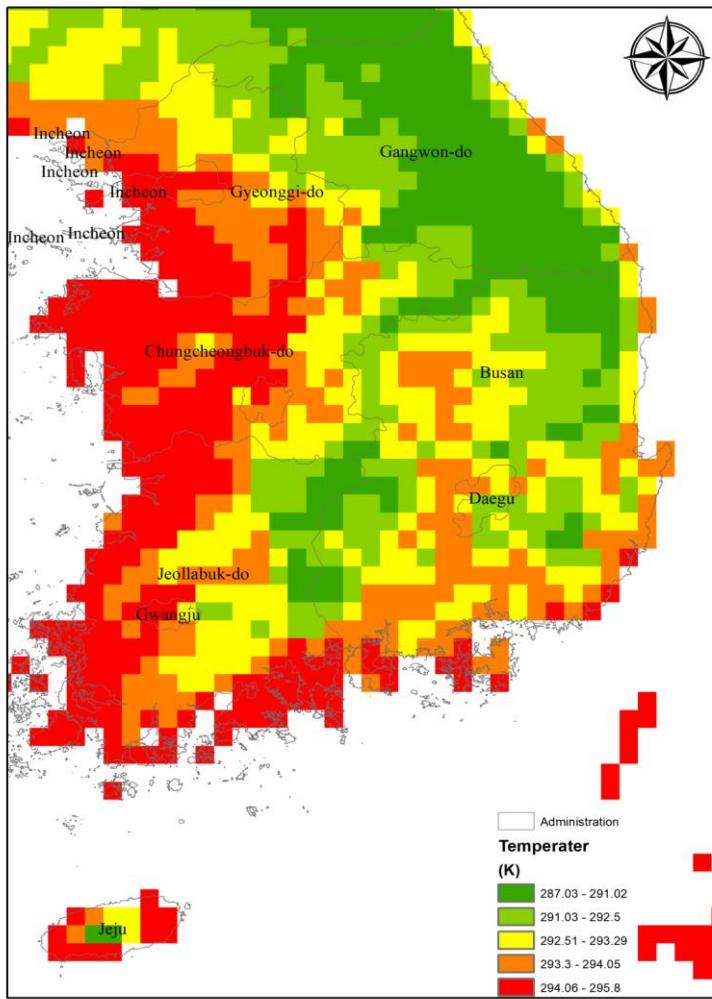


Fig 24.Spatial temperature coverages on September 2015

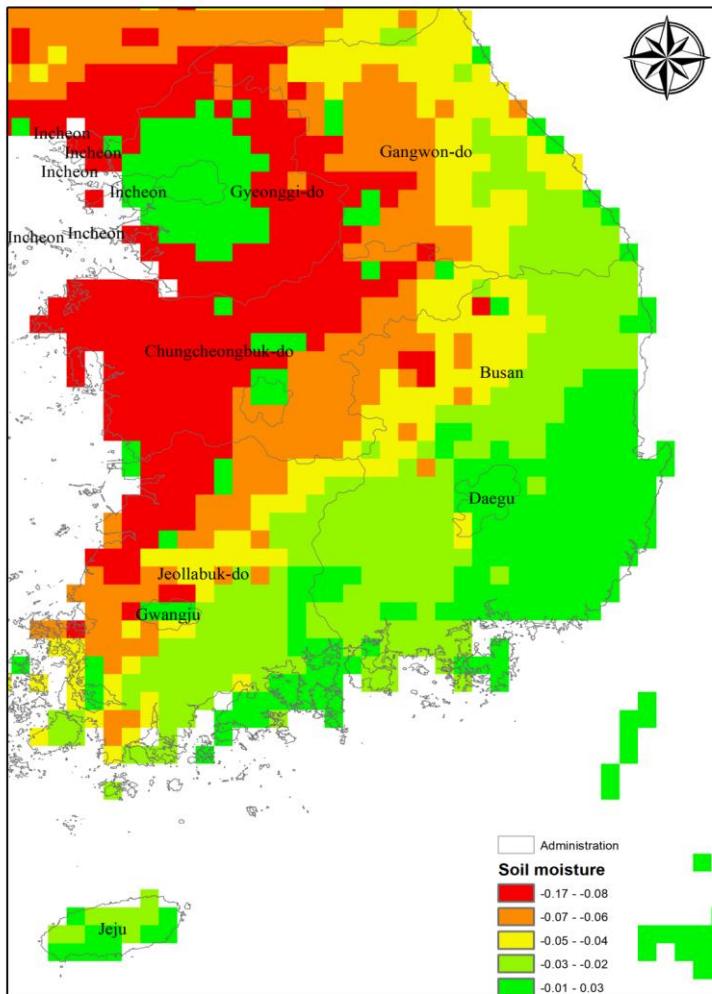


Fig 25. Spatial soil moisture coverages on September 2015

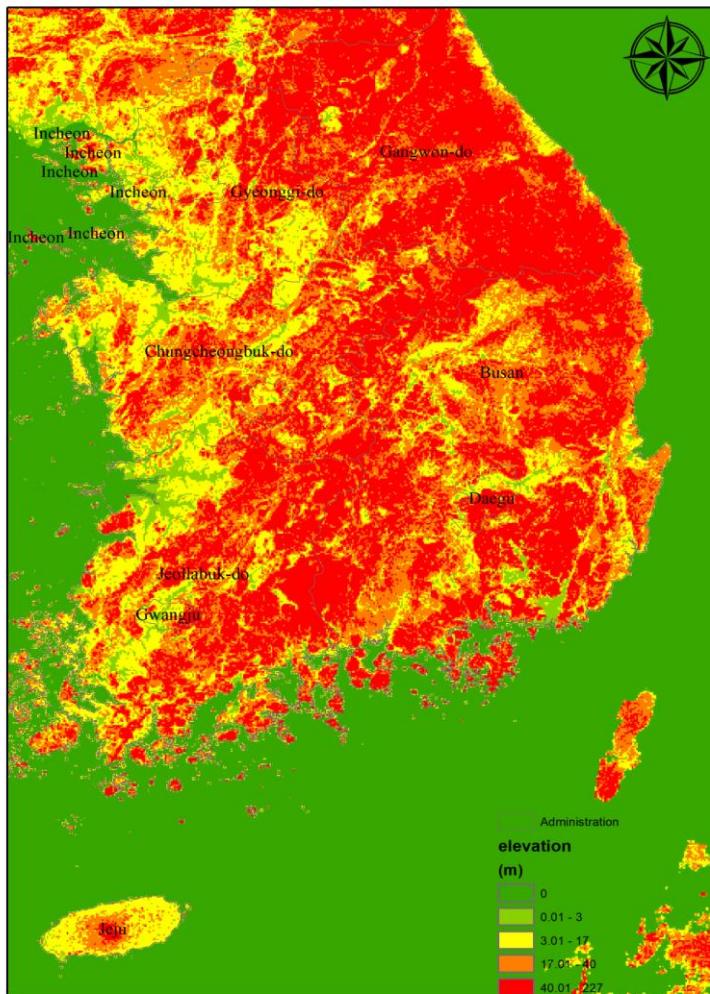


Fig 26. Topography of the study area

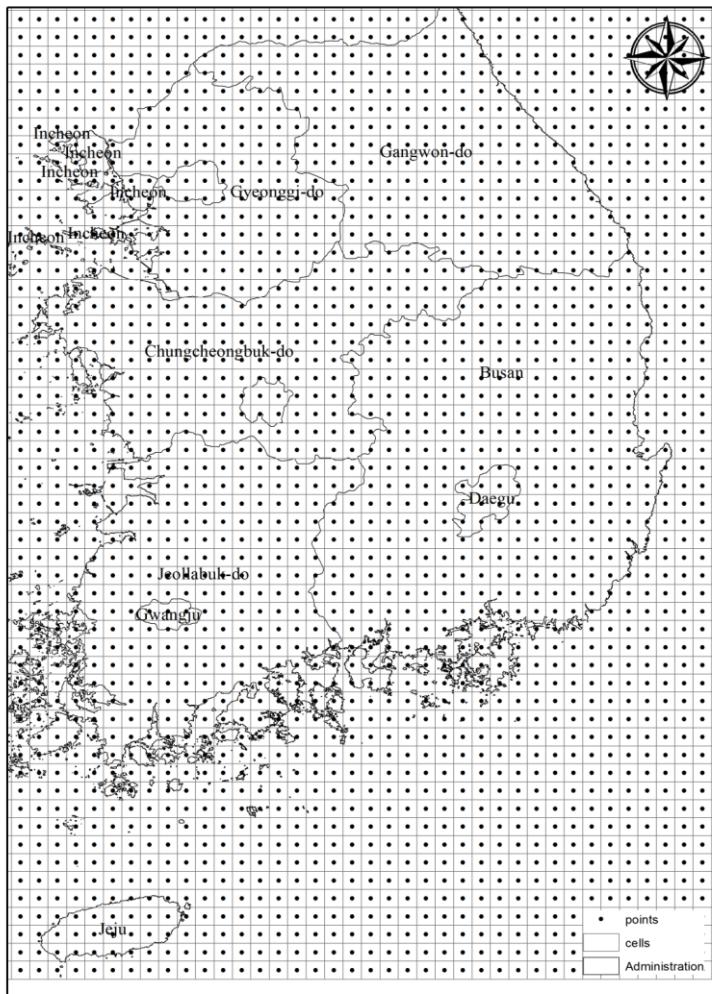


Fig 27. Grid cells of the study

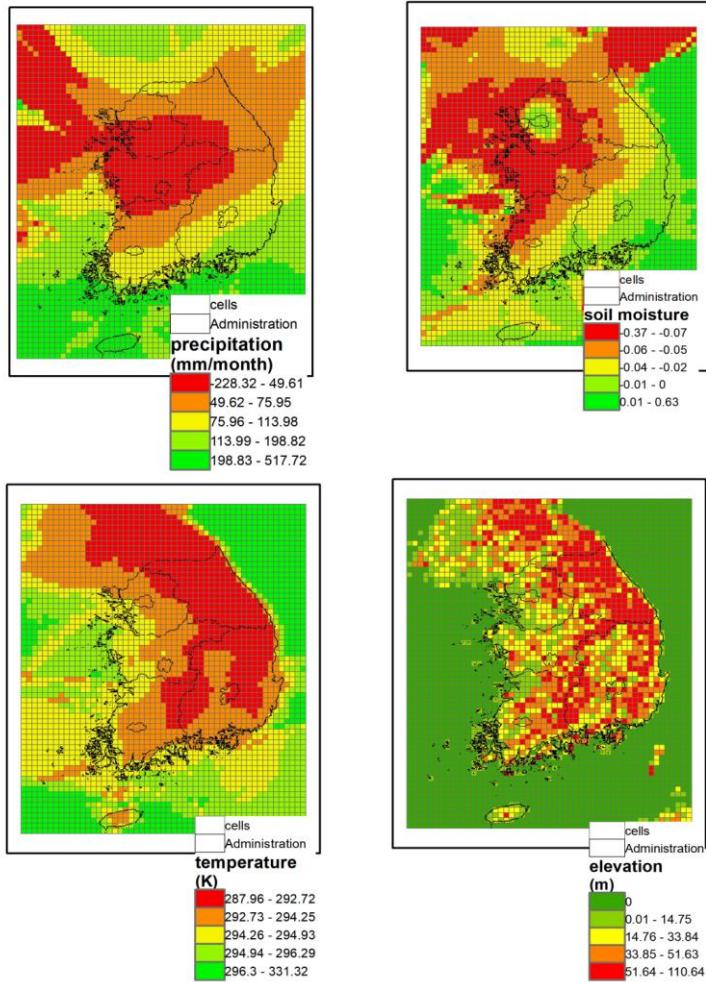


Fig 28. The satellite-based data were transformed to grid cells

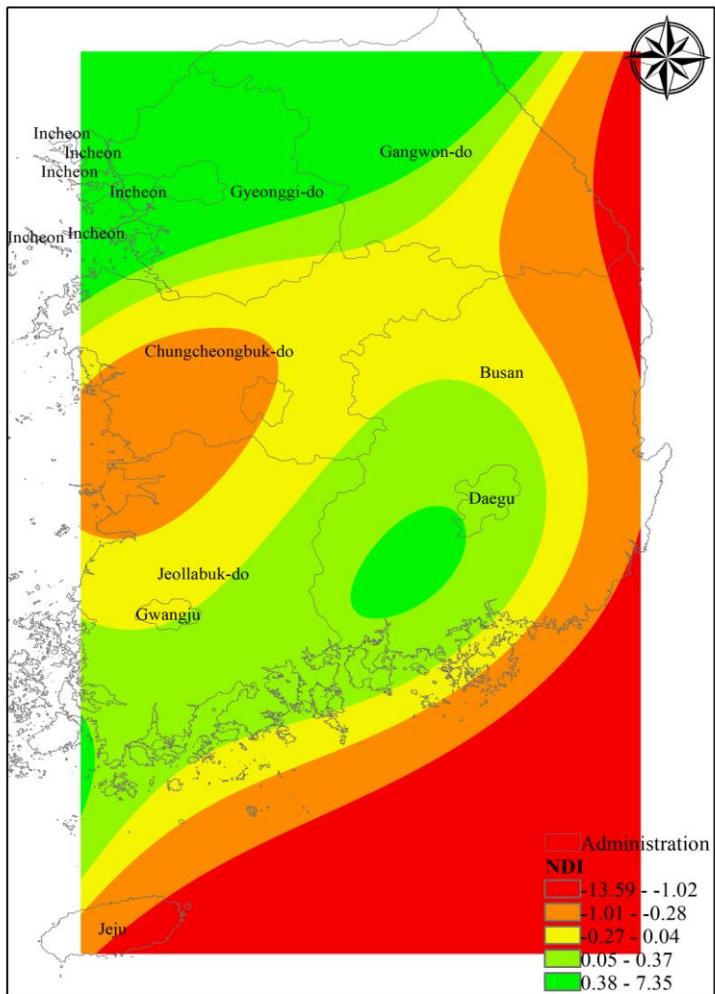


Fig 29. Spatial trend of NDI

**Professor comments:**

- He is angry because we think 59 ASOS is not enough. He informs that this is official data in South Korea. We should not talk to him about data again.
- The extrapolation is a technique skill. He does not care about it. We should handle it by ourself.
- He does not understand what is main objective of our study
- We should complete it first. Then I will give his comments.

## WEEK22

A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network

### 6. Data processing (continue)

*Table 5. The status of processing data analysis*

	Contents	Sub-contents
Week 1-2	<input checked="" type="checkbox"/> Data processing	21. Convert NDI to spatial-space format 22. Compute monthly average 23. Determine the specific time of the maximum value NDI 24. Retrieved precipitation from satellite 25. Retrieved temperature from satellite 26. Retrieved soil moisture from satellite 27. Retrieved elevation from satellite 28. Covert satellite-based data to grid-point and grid cell format 29. Analysis spatial trend of NDI
Week 3	<input type="checkbox"/> ANN model	30. Create surfaces for NDI, satellite-based data, longitude, latitude follow the spatial resolutions: 0.1 degree 31. Extract satellite-based precipitation, soil moisture, temperature, elevation, runoff, evapotranspiration, longitude, latitude as 59 ASOS location. Check correlation of NDI and auxiliary variables 32. Set up the neural network regression (NNR) to predict NDI from auxiliary variables above 33. Build the structure of ANN 34. Training and validate ANN 35. Optimize the structure of ANN 36. Create a table to evaluate the model
Week 4	<input type="checkbox"/> Merging-condition merging model	37. Compute the residual 1 of ANN 38. Compute residual 2 of Kriging regression 39. Compute Cokriging-conditional merging variogram of NN and Kriging 40. Analysis spatial NDI coverage by using conditional variogram

We extracted values of NDI and auxiliary at each ASOS location in step 11. To figure out the relationship of them, the multiple corellated analysis was used. The correlaltion are shown in Fig 1.

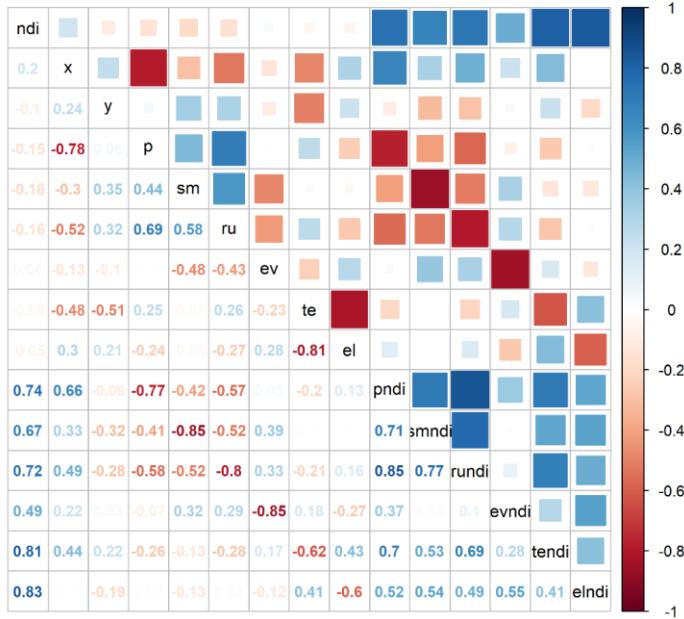


Fig 30. Correlation of ndi, x( longitude), y(lattitue), p( precipitation), sm( soil moisture), ev (evapotranspiration), te (temperature), el( elevation), pndi (bias of precipitation- ndi), smndi( bias of soil moisture-ndi), rundi (bias of runoff – ndi), evndi (bias of evapotranspiration-ndi),tendi( bias of temperature-ndi), elndi (bias of elevation-ndi).

The result show that ndi has a strong correlltion to runoff (0.58), elevation (0.81), bias of soil moisture- ndi (0.71) , bias of runoff -ndi (0.77). The rest of them has a correlation below to 0.5.

The satellite-based precipitation and the bias of precipitation- ndi have not a strong correlation. It shows the opposite result to the initial computation. Therefore, we propose to check the processing of collection data before setup the regression model of ANN. Because the chosen of auxiliary variable can impact to the accuracy of the model.

## 7. Reading the commented papers

NO	purpose	Method	result	References
1	Assess agricultural drought using remote sensing approach	<ul style="list-style-type: none"> <li>Using remote sensing data from NOAA-AVHRR NDVI to computed drought by SPI, NDVI, VCI. Analys the correlliton of them.</li> </ul>	<ul style="list-style-type: none"> <li>✓ The correlation coefficient between VCI and yield of major rain-fed crops (<math>r &gt; 0.75</math>) also supports the efficiency of this remote sensing derived index for assessing agricultural drought</li> </ul>	Dutta et al 2015
2	Compare meteorological drought by using SPI, RDI	<ul style="list-style-type: none"> <li>Using historical precipitation for 40 years for statiscal analaysis</li> </ul>	<ul style="list-style-type: none"> <li>✓ The correlation between the SPI and RDI indices is good in the different time scales confirming their performance.</li> <li>✓ The RDI gives the greater number of drought months, but both methods show the same drought durations and severity as the results of the Deciles index</li> </ul>	Haied al, 2017
3	Develop a package for drought monitoring, prediction and assessment	<ul style="list-style-type: none"> <li>Develop program in R languges</li> </ul>	<ul style="list-style-type: none"> <li>✓ Computation drought index included univariate drought index and multivariate drought index</li> <li>✓ Joint distribution and Kendall distribution function</li> <li>✓ Statistical drought prediction included based line and multivariate drought prediction</li> </ul>	Hao et al, 2017

NO	purpose	Method	result	References
4	Analysis the global database of meteorological drought events from 1951 to 2016	<ul style="list-style-type: none"> <li>three spatial scales: global (<math>0.5^\circ</math>), macro-regional, and country scale was studied. The data base of drought events were collected from the Global Drought Observatory of the European Commission's Joint Research Centre.</li> <li>Events were detected at macro-regional and country scale based on the separate analysis of the SPEI and SPI at different accumulation scales (from 3 to 72 months)</li> <li>Using as input the Global Precipitation Climatology Centre (GPCC) and Climatic Research Unit (CRU) Time Series datasets.</li> </ul>	<p>✓ The database includes approximately 4800 events based on SPEI-3 and 4500 based on SPI-3. Each event is described by its start and end date, duration, intensity, severity, peak, average and maximum area in drought, and a special score to classify 52 mega-droughts.</p>	Spinoni et al 2019

#### Extracted idea from our articles:

Remote sensing is one of the usefull resource for drought analysis. We can exploit it to estimate drought. Comparison various drought indices is also the interesting. The correlation of different drought indice can be used as the metric for evaluation of model. The global drought events can be used as the references when we overview the the drought problem.

#### References

- Dutta, D., Kundu, A., Patel, N. R., Saha, S. K., & Siddiqui, A. R. (2015). Assessment of agricultural drought in Rajasthan (India) using remote sensing derived Vegetation Condition Index (VCI) and Standardized Precipitation Index (SPI). *The Egyptian Journal of Remote Sensing and Space Science*, 18(1), 53-63. doi:<https://doi.org/10.1016/j.ejrs.2015.03.006>
- Haied, N., Foufou, A., Chaab, S., Azlaoui, M., Khadri, S., Benzahia, K., & Benzahia, I. (2017). Drought assessment and monitoring using meteorological indices in a semi-arid region. *Energy Procedia*, 119, 518-529. doi:<https://doi.org/10.1016/j.egypro.2017.07.064>
- Hao, Z., Hao, F., Singh, V. P., Ouyang, W., & Cheng, H. (2017). An integrated package for drought monitoring, prediction and analysis to aid drought modeling and assessment. *Environmental Modelling & Software*, 91, 199-209. doi:<https://doi.org/10.1016/j.envsoft.2017.02.008>

Spinoni, J., Barbosa, P., De Jager, A., McCormick, N., Naumann, G., Vogt, J. V., . . . Mazzeschi, M. (2019). A new global database of meteorological drought events from 1951 to 2016. *Journal of Hydrology: Regional Studies*, 22, 100593. doi:<https://doi.org/10.1016/j.ejrh.2019.100593>

**Professor comments:**

- Correct fig 4
- Be careful to use the name of obs. We have not observational data
- The scale of both x and y as the same



## WEEK 23

A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network

### 8. Data processing (continue)

*Table 6. The status of processing data analysis*

	Contents	Sub-contents
Week 1-2	<input checked="" type="checkbox"/> Data processing	<ul style="list-style-type: none"> <li>①. Convert NDI to spatial-space format</li> <li>②. Compute monthly average</li> <li>③. Determine the specific time of the maximum value NDI</li> <li>④. Retrieved precipitation from satellite</li> <li>⑤. Retrieved temperature from satellite</li> <li>⑥. Retrieved soil moisture from satellite</li> <li>⑦. Retrieved elevation from satellite</li> <li>⑧. Covert satellite-based data to grid-point and grid cell format</li> <li>⑨. Analysis spatial trend of NDI</li> </ul>
Week 3	<input type="checkbox"/> ANN model	<ul style="list-style-type: none"> <li>⑩. Create surfaces for NDI, satellite-based data, longitude, latitude follow the spatial resolutions: 0.1 degree</li> <li>⑪. Extract satellite-based precipitation, soil moisture, temperature, elevation, runoff, evapotranspiration, longitude, latitude as 59 ASOS location. Check correlation of NDI and auxiliary variables</li> <li>⑫. Set up the neural network regression (NNR) to predict NDI from auxiliary variables above</li> <li>⑬. Build the structure of ANN</li> <li>⑭. Training and validate ANN</li> <li>⑮. Optimize the structure of ANN</li> <li>⑯. Create a table to evaluate the model</li> </ul>
Week 4	<input type="checkbox"/> Merging-conditional model	<ul style="list-style-type: none"> <li>⑰. Compute the residual 1 of ANN</li> <li>⑱. Compute residual 2 of Kriging regression</li> <li>⑲. Compute Cokriging-conditional merging variogram of NN and Kriging</li> <li>⑳. Analysis spatial NDI coverage by using conditional variogram</li> </ul>

### 3.9. Set up the neural network regression (NNR) to predict NDI from auxiliary variables

After checking the retrieved processing data from satellite, we limited the spatial boundary of our study is from 125.80 to 129.70 degree of longitude, and from 33.129 to 38.70 degree of latitude. Boundaril study area was adjusted to remove irrelevant data. Data was normalized to set the nondimensional unit and set the same scale from 0 to 1. Overview data is presented by the graphic of boxplot in Fig 1 below :

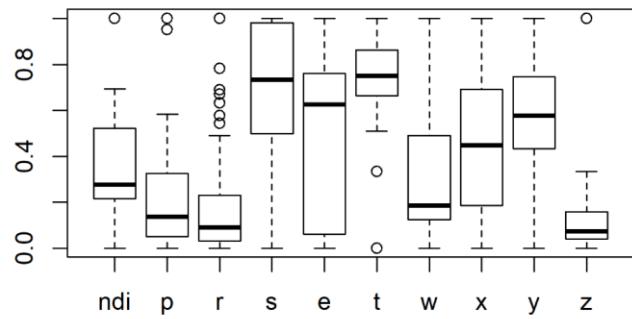


Fig 31. The normalized data of ndi, p (precipitation), r (runoff), s (soil moisture), e (evapotranspiration), t (temperature), w (wind), x (latitude), y (longitude), z (elevation).

Fig 2 shows that ndi is significantly correlate to soil moisture(s), runoff (r) and evapospiration (e).

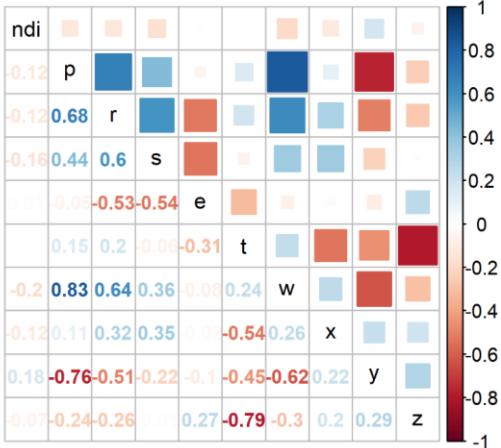


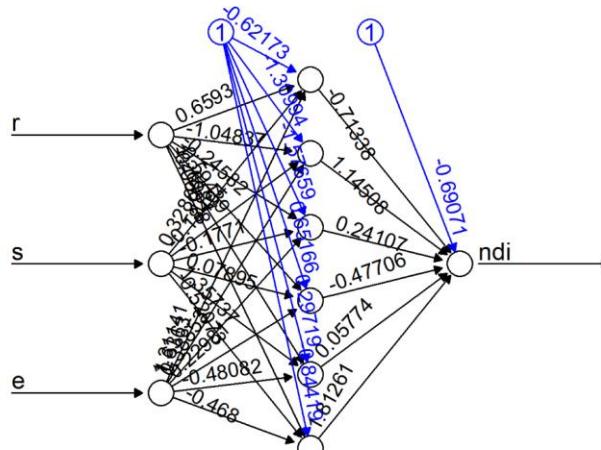
Fig 32. The correlation of dataset

### 3.10. Build the structure of ANN

The first trial artificial neural network (ANN) has 1 one layer, and consist of 10 neurons is used as the initial ANN structure. We use a customized function softplus as the active function follow:

$$f = \log(1 + e^x) \quad (1)$$

Various active function was examined in further steps. The Fig 3 shows the structure of ANN. Ndi is determined by summarize of these multiple r, s, e with 6 neuron by weight number.



Error: 1.264534 Steps: 58

Fig 33. Structure of ANN

### 3.11. Training and validate ANN

The dataset is divided into training data set and testing data set follow the ratio 0.7 and 0.3. We also random split dataset into 10 subsample data.

The result of using ANN to predict ndi values based on runoff, soil moisture and evapotranspiration is very poor because it was not optimized.

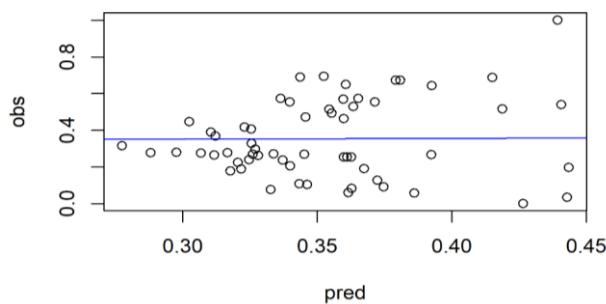


Fig 34. Observation and predict value using ANN

## 9. Discussion

We propose optimizing ANN follow by :

- (1) Changing the number of neuron and number of hidden layers.
- (2) Changing active function
- (3) Changing the number of auxiliary variable in neural network regression. It means that we will consider also the small correlation variable such as precipitation, temperature, wind speed, elevation. We are expected that ANN learn from non-linearity.
- (4) Expanding data training and testing from 2001 to 2016. We will exploit data as much as possible. Instead of using extreme period (September 2015), we will use whole available data for setup model.

## WEEK 24

**A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network**

### 10. ANN model

*Table 7. The status of processing data analysis*

	Contents	Sub-contents
Week 1-2	<input checked="" type="checkbox"/> Data processing	<ul style="list-style-type: none"> <li>①. Convert NDI to spatial-space format</li> <li>②. Compute monthly average</li> <li>③. Determine the specific time of the maximum value NDI</li> <li>④. Retrieved precipitation from satellite</li> <li>⑤. Retrieved temperature from satellite</li> <li>⑥. Retrieved soil moisture from satellite</li> <li>⑦. Retrieved elevation from satellite</li> <li>⑧. Covert satellite-based data to grid-point and grid cell format</li> <li>⑨. Analysis spatial trend of NDI</li> </ul>
Week 3	<input type="checkbox"/> ANN model	<ul style="list-style-type: none"> <li>⑩. Create surfaces for NDI, satellite-based data, longitude, latitude follow the spatial resolutions: 0.1 degree</li> <li>⑪. Extract satellite-based precipitation, soil moisture, temperature, elevation, runoff, evapotranspiration, longitude, latitude as 59 ASOS location. Check correlation of NDI and auxiliary variables</li> <li>⑫. Set up the neural network regression (NNR) to predict NDI from auxiliary variables above</li> <li>⑬. Build the structure of ANN</li> <li>⑭. Training and validate ANN</li> <li>⑮. Optimize the structure of ANN</li> <li>⑯. Create a table to evaluate the model</li> </ul>
Week 4	<input type="checkbox"/> Merging-conditional model	<ul style="list-style-type: none"> <li>⑰. Compute the residual 1 of ANN</li> <li>⑱. Compute residual 2 of Kriging regression</li> <li>⑲. Compute Cokriging-conditional merging variogram of NN and Kriging</li> <li>⑳. Analysis spatial NDI coverage by using conditional variogram</li> </ul>

### 3.1. Optimize the structure of ANN

#### (1) Changing the number of neuron and number of hidden layers.

We estimate the impact of number neurons to the accuracy of model by change number of neurons.

The number of neurons was changed from 1 to 10. We estimate the maen absolute error (MAE) of computed ndi values and ANN predicted ndi value. Fig1 shows 1 layers and 7 neurons give the best prediction (the lowest value of MAE).

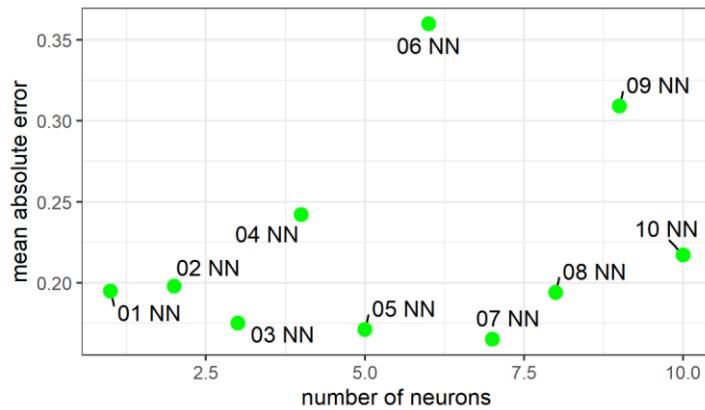


Fig 35. Number of neurons impacts to the accuracy of the model

#### (2) Changing active function

We fixed the ANN model using 1 layer, 7 neurons, and change active functions to examine impact of active function. Seven active functions (tanh, logistic, softplus, relu, sigmoid, leakyrelu, swish) are scrutinized. Using 7 neurons, the model is unstable with leakyrelu active function. Therefore, we use 5 neurons for testing all active functions. The swish active function gives the best results with the lowest value of mae (Fig 2).

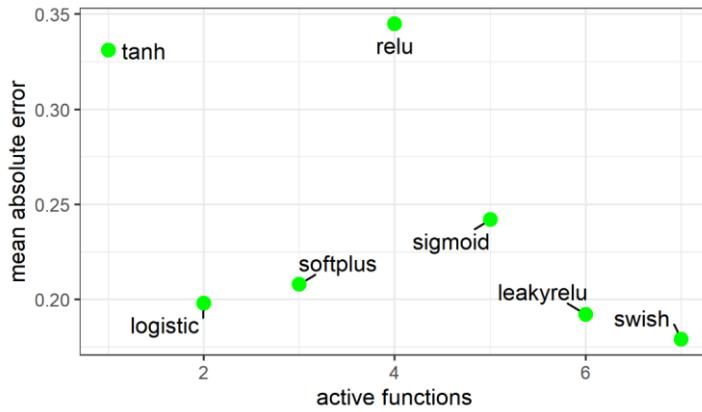


Fig 36. Active function impacts to the accuracy of ANN model

### (3) Changing the number of auxiliary variable in the neural network regression

We choose 1 layer, 5 neurons, swish active function as a baseline model. Then, the neural network regression model was analyzed with diversified auxiliary variables. The auxiliary variables were chosen from 1 to 9 (precipitation, runoff, soil moisture, evapotranspiration, temperature, wind speed, longitude, latitude, topology elevation). The result shows that the combination of precipitation (p), runoff (r), soil moisture (s), evapotranspiration (e) and temperature (t) give the lowest mean absolute value (Fig 3).

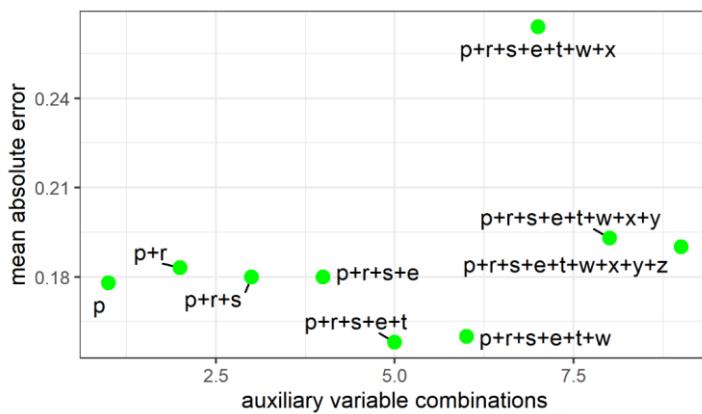


Fig 37. Combination of auxiliary variables impacts to the accuracy of the model

The comparison of computing ndi and ANN-based ndi was presented in Fig 4.

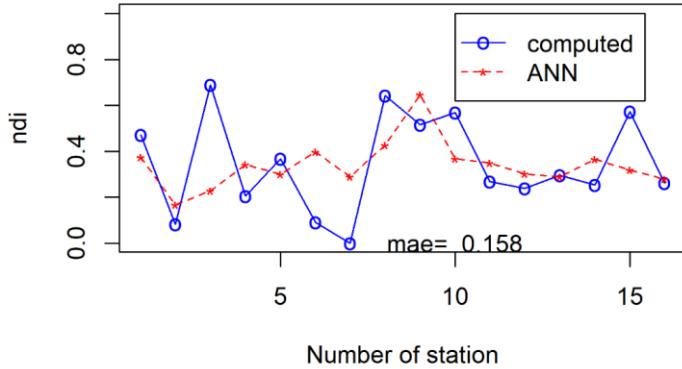
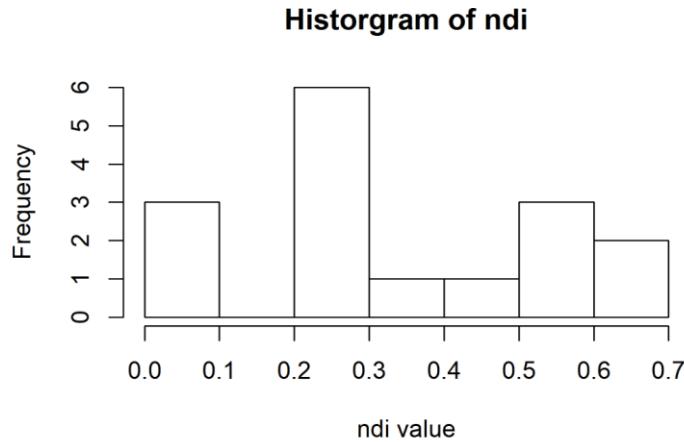


Fig 38. Comparison of computed ndi and ANN-based ndi using 1 layer, swish active function and combination of  $p$ ,  $r$ ,  $s$ ,  $e$ ,  $t$  as predictors.

At the first and the second values of testing, ANN-based ndi is approximately matched to computed ndi. In the third value of testing ANN-based has the increased trend fit to computed ndi. But the fourth value of testing is opposite to the computed ndi. Similarly, the 5th, the 6<sup>th</sup>, 7<sup>th</sup>, 9<sup>th</sup>, 10<sup>th</sup>, 13<sup>th</sup>, 14<sup>th</sup>, 15<sup>th</sup> of testing values are insufficient results. Fig 5 present historical gram of testing data set.



*Fig 39. Historygram of ndi testing data set*

From Fig 5 and Fig 4, we found the ratio of the mean absolute error to range ndi value (0.158/0.700~0.225) is approximate 22.5%. It is not so bad result.

### 3.2. Create a table to evaluate the model

#### Conclusion and discussion

(1) The ANN model was optimized by comparing of 26 models (10 models changing number of neurons, 7 models changing active functions, 9 models changing auxiliary variable). The model uses 5 neurons, swish active function, p, r, s, e, t as predictors is the most suitable. It gives the lowest value of absolute mean error for testing data. The error rate is 22.5 %. This result is obtained by sequence analysis. It means that we separately estimate the impact of number neurons, active function, auxiliary variables. The interaction with them was not considered. The random combination of 26 models should be figured out in the future.

(2) The model using the limited training data included the ndi in Septemer 2015 at 43 for training and 16 stations for testing. Totally,  $43 \times 10 = 430$  values was used for training, and  $16 \times 10 = 160$  values was used for testing. The almost satellite-based data is available from 2001. If we extended

period analysis from 1 months (September 2015) to 192 months (2001-2016), the accuracy of model could be improved. However, it will include the non-extreme drought events and make the model is more complicated.

(3) The other approach to improve model that we using multiple layer of neuron (deep learning) algorithm. By increasing the number of layers, optimize parameters (loss functions, gradient descent hyperparameters, learning rate, batch size, iterative update, cost functions). Compare to extend data, this approach is inspite of the time-consuming, but it helps us understand deeply about machine learning. It is useful for futural studies. Therefore, we propose to continue optimizing ANN by using a deep neural network approach in the future

(4) In the next step, we propose to continue computing of merging conditional model (residual of ANN and residual of Original Kringing). Build the conditional variogram and compare to isolated variograph of ANN model, OK model. The expected result of OK-ANN will be better than separate model. This is the main objective of our study.

## Appendix

List of figures compare ANN-based ndi and computed ndi

21. Impact of number neuron

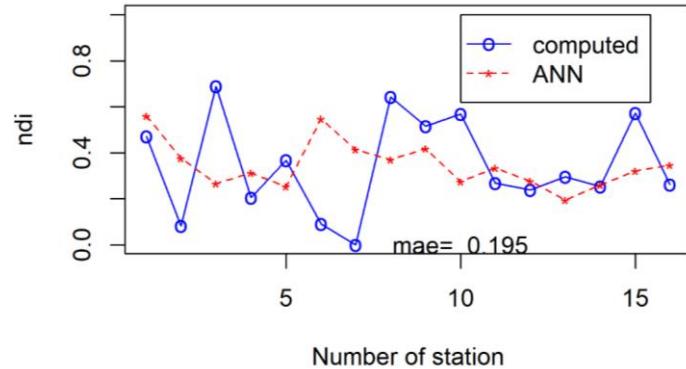


Fig 40. Using 1 neuron

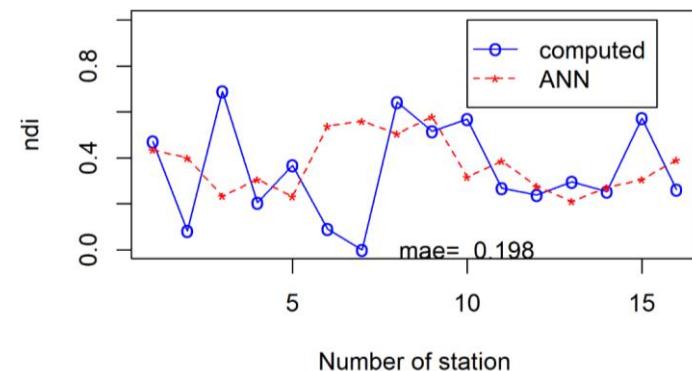


Fig 41. Using 2 neurons

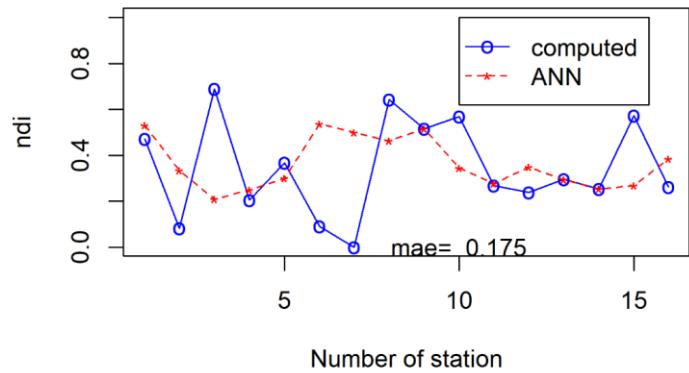


Fig 42. Using 3 neurons

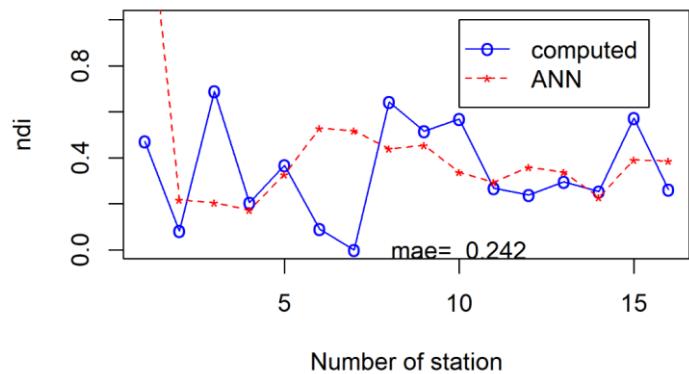


Fig 43. Using 4 neurons

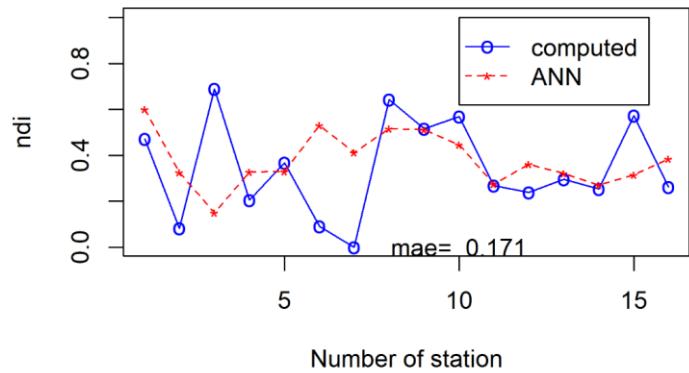


Fig 44. Using 5 neurons

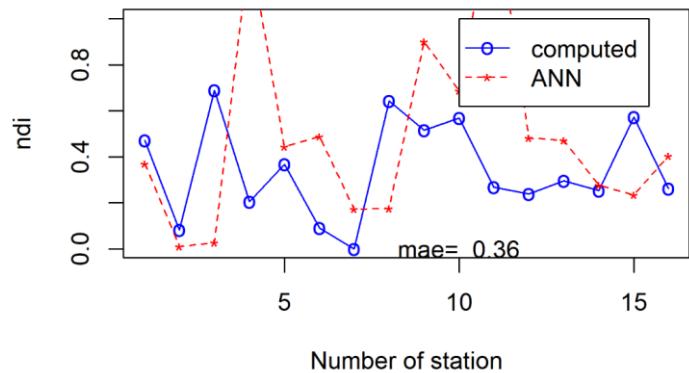


Fig 45. Using 6 neurons

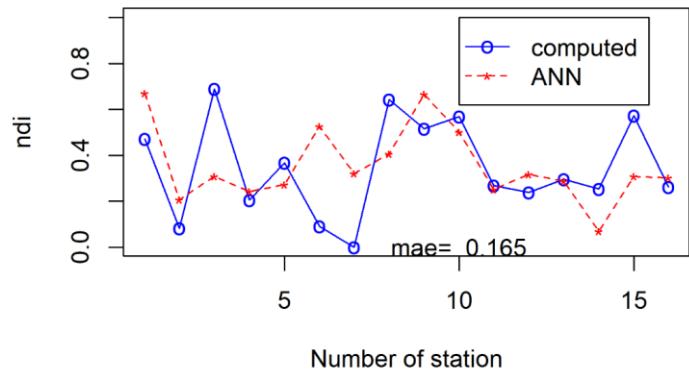


Fig 46. Using 7 neurons

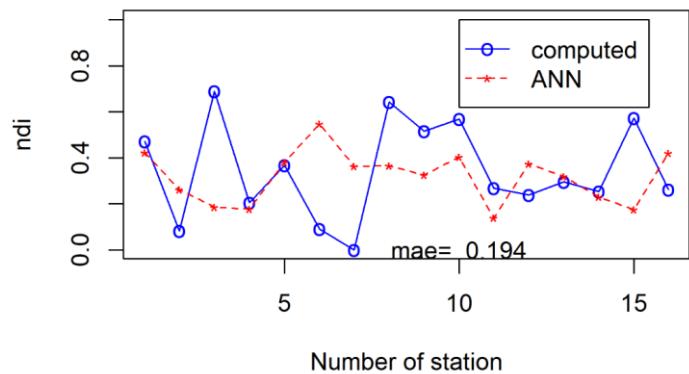


Fig 47. Using 8 neurons

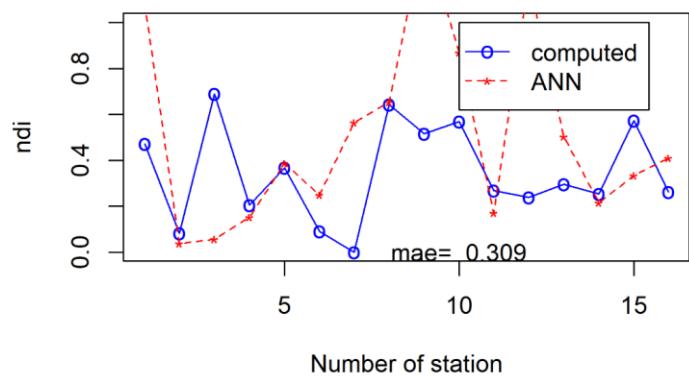


Fig 48. Using 9 neurons

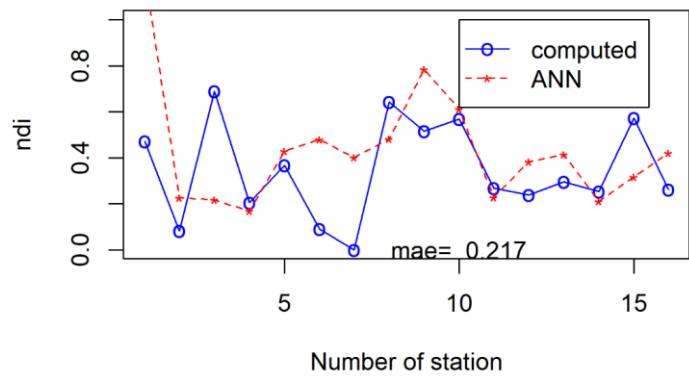


Fig 49. Using 10 neurons

## 22. Impact of active functions

Fig 50. Using tanh active function

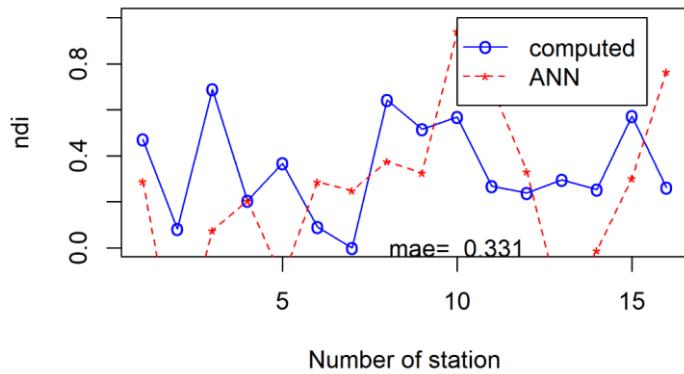


Fig 51. Using tanh active function

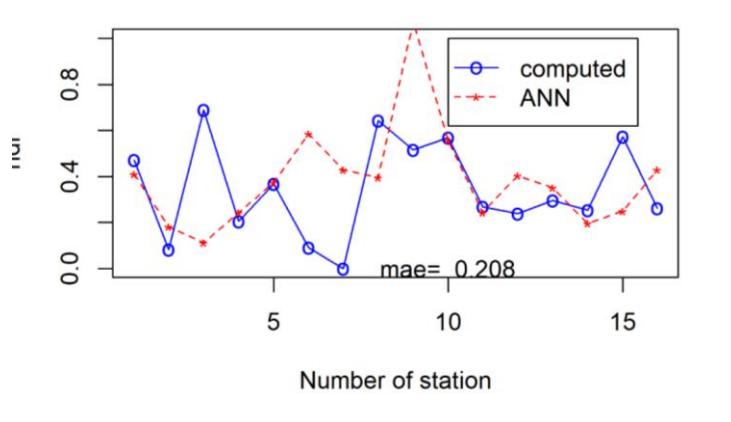


Fig 52. Using softplus active function

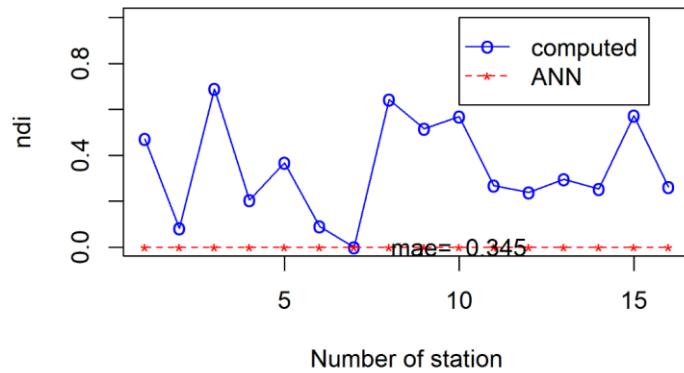


Fig 53. Using relu active function

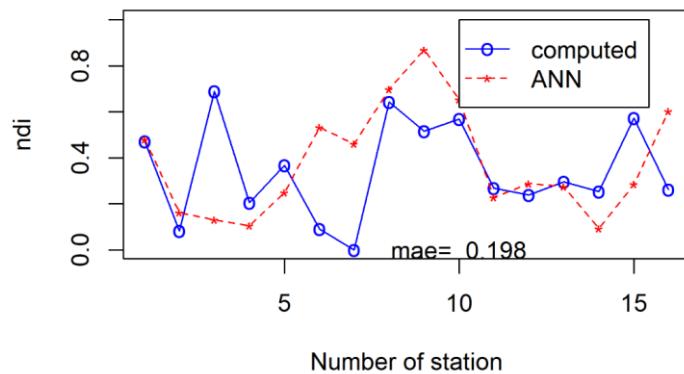


Fig 54. Using logistic active function

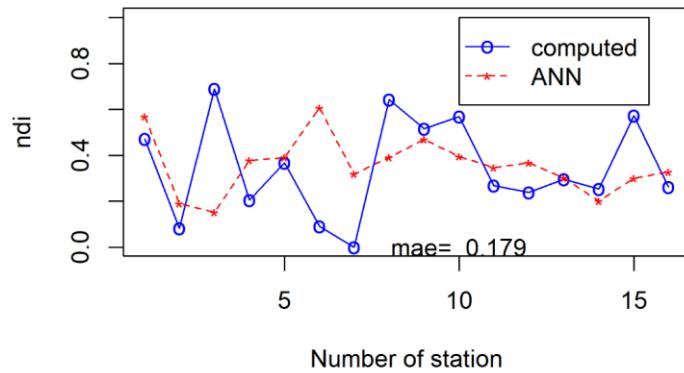


Fig 55. Using swish active function

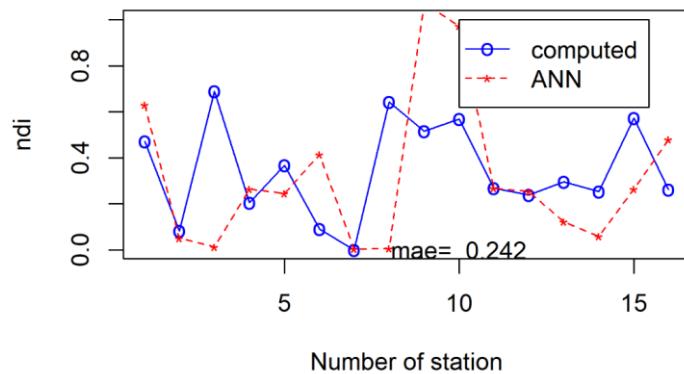


Fig 56. Using sigmoid active function

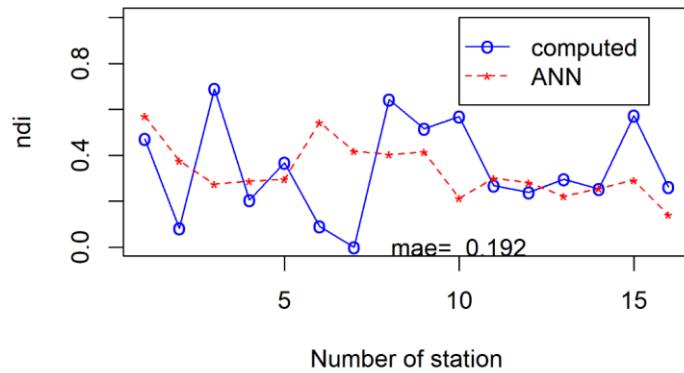


Fig 57. Using leakyrelu active function

23. Impact of auxiliary variables

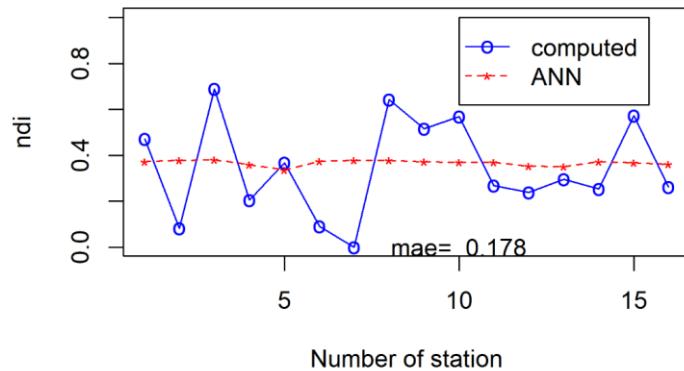


Fig 58. Using only precipitation

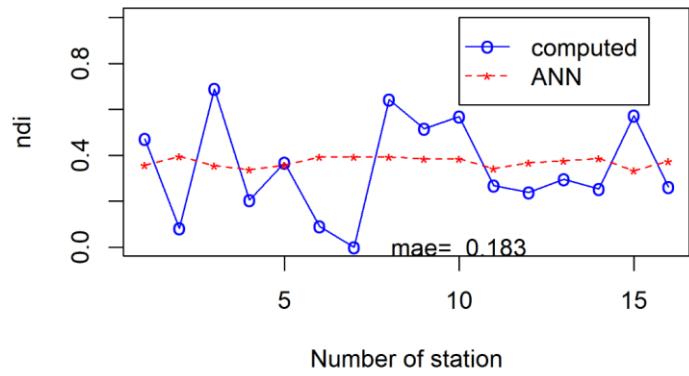


Fig 59. Using precipitation and runoff

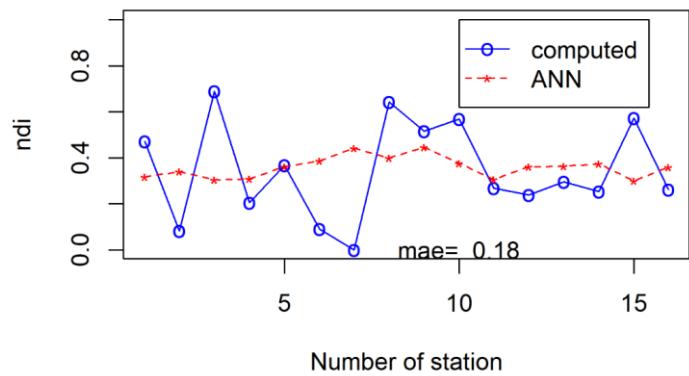


Fig 60. Using precipitation, runoff and soil moisture

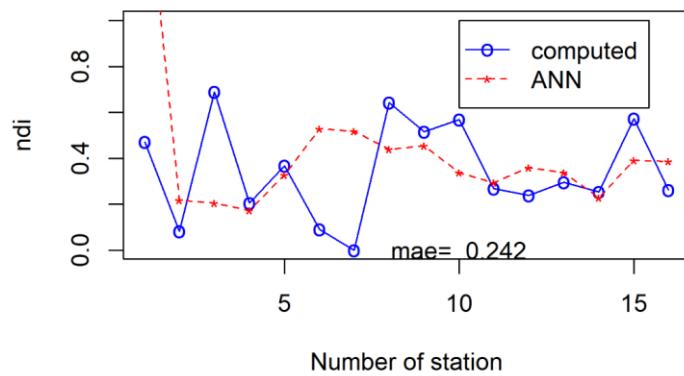


Fig 61. Using precipitation, runoff, soil moisture and evapotranspiration

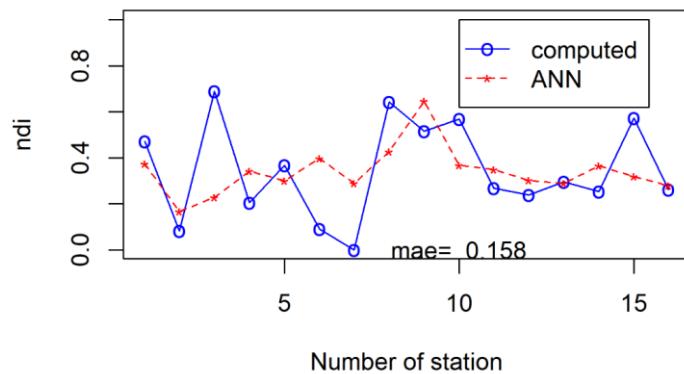


Fig 62. Using precipitation, runoff, soil moisture, evapotranspiration and temperature

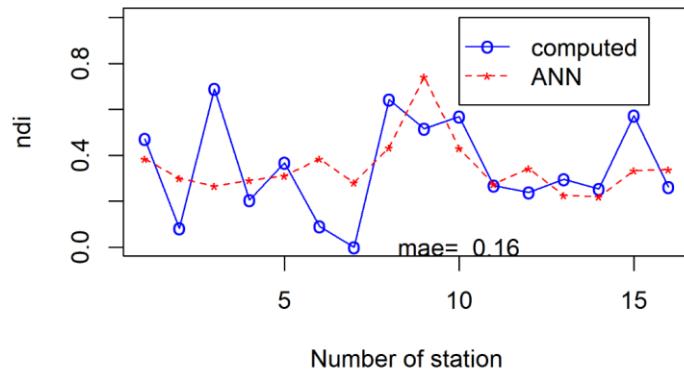


Fig 63. Using precipitation, runoff, soil moisture, evapotranspiration, temperature and wind

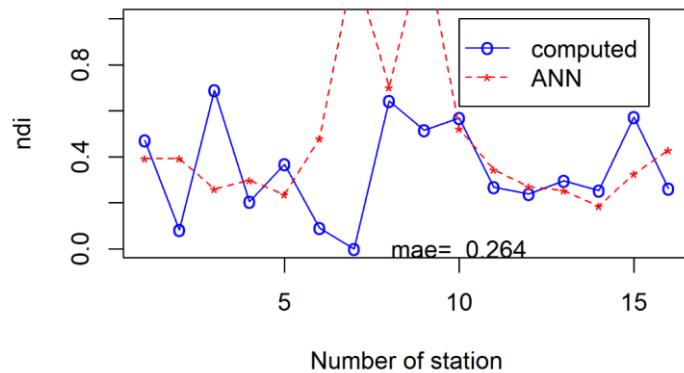


Fig 64. Using precipitation, runoff, soil moisture, evapotranspiration, temperature, wind and longitude

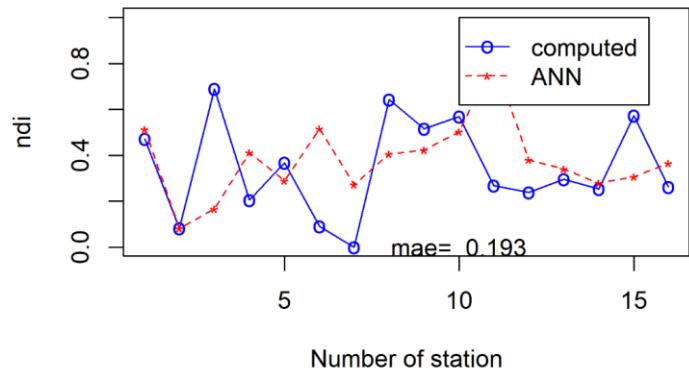


Fig 65. Using precipitation, runoff, soil moisture, evapotranspiration, temperature, wind, longitude and latitude

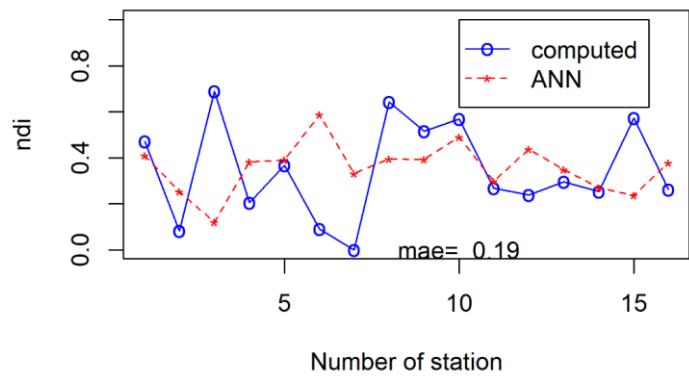


Fig 66. Using precipitation, runoff, soil moisture, evapotranspiration, temperature, wind, longitude, latitude and topological elevation

## WEEK 25

**A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network**

### 11. Merging conditional model

*Table 8. The status of processing data analysis*

	Contents	Sub-contents
Week 1-2	<input checked="" type="checkbox"/> Data processing	<ul style="list-style-type: none"><li>①. Convert NDI to spatial-space format</li><li>②. Compute monthly average</li><li>③. Determine the specific time of the maximum value NDI</li><li>④. Retrieved precipitation from satellite</li><li>⑤. Retrieved temperature from satellite</li><li>⑥. Retrieved soil moisture from satellite</li><li>⑦. Retrieved elevation from satellite</li><li>⑧. Covert satellite-based data to grid-point and grid cell format</li><li>⑨. Analysis spatial trend of NDI</li></ul>
Week 3	<input type="checkbox"/> ANN model	<ul style="list-style-type: none"><li>⑩. Create surfaces for NDI, satellite-based data, longitude, latitude follow the spatial resolutions: 0.1 degree</li><li>⑪. Extract satellite-based precipitation, soil moisture, temperature, elevation, runoff, evapotranslation, longitude, latitude as 59 ASOS location. Check correlation of NDI and auxiliary variables</li><li>⑫. Set up the neural network regression (NNR) to predict NDI from auxiliary variables above</li><li>⑬. Build the structure of ANN</li><li>⑭. Training and validate ANN</li><li>⑮. Optimize the structure of ANN</li><li>⑯. Create a table to evaluate the model</li></ul>

	Contents	Sub-contents
Week 4	<input type="checkbox"/> Merging-conditional model	<p>(17). Compute the residual 1 of ANN  (18). Compute residual 2 of Kriging regression  (19). Compute Corking-conditional merging variogram of NN and Kriging  (20). Analysis spatial NDI coverage by using conditional variogram</p>

### 3.12. Compute the residual 1 of Kriging

The residual of Kriging shows the difference between of NDI-based on computation and NDI-based on kriging model prediction. Using NDI-based on Kriging prediction, we determined the variogram. The variogram was computed follow studies (Pebesma, 2004; Pebesma & Wesseling, 1998). The variogram is present in the Fig 1.

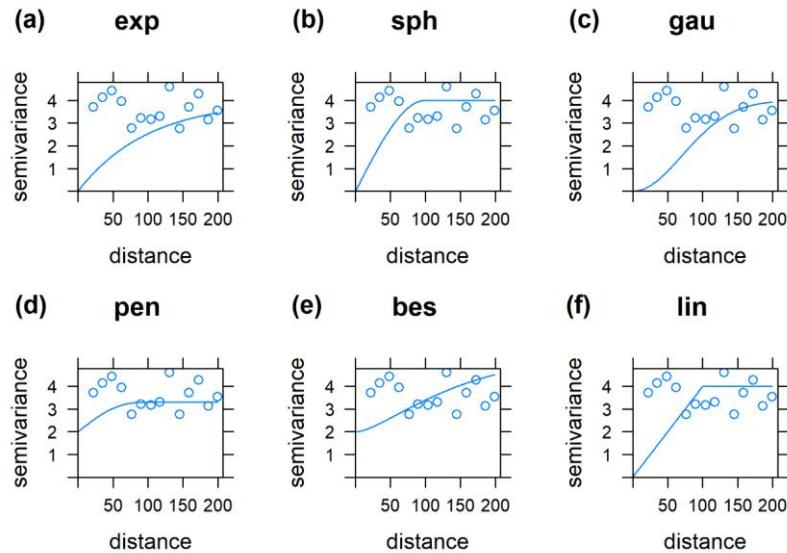


Fig 67. Variogram of NDI using: (a) Exponential, (b) Spherical, (c) Gaussian, (d) Pentaspherical, (e) Bessel, (f) Linear model

To quantify the best fit of the model, we estimate the sum square error. Results are presented in table 2.

Table 9. Error of variogram using various models with Kriging

No.	Model	Sum square error
1	Pentaspherical	0.0391
2	Spherical	0.0408
3	Linear	0.0421
4	Exponential	0.0478
5	Gaussian	0.0507
6	Bessel	0.0607

Results show that Pentaspherical model is the most fit for variogram with the lowest sum square error.

$$SSE = \sum_i^n (x_i - \bar{x})^2 \quad (1)$$

Where n is the number of observations,  $x_i$  is the value of the  $i$ th observation and  $\bar{x}$  is the mean of all the observations. The smallest value of the error sum of squares (SSE) denotes for better result. SSE is the sum of the squared differences between each observation and its group's mean. It can be used as a measure of variation within a cluster. If all cases within a cluster are identical the SSE would then be equal to 0.

### 3.13. Compute the residual 2 of ANN

The residual of NDI-based on computation and NDI-based on Kriging was present in Table 3. Results show that Bessel model is the best fit for variogram of ANN.

Table 10. Error of variogram using various models with ANN

No.	Model	Sum square error
1	Bessel	0.0017
2	Exponential	0.0021
3	Gaussian	0.0095
4	Linear	0.0019
5	Pentaspherical	0.0033

6 Spherical	0.0028
-------------	--------

Various variograms models using ANN are presented in Fig 2

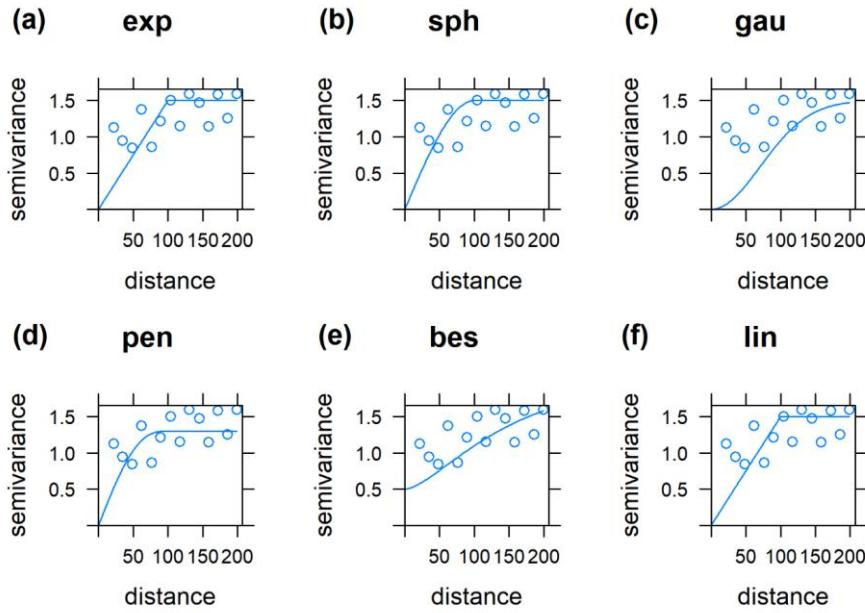


Fig 68. Variogram of NDI using: (a) Exponential, (b) Spherical, (c) Gaussian, (d) Pentaspherical, (e) Bessel, (f) Linear model

Compare to Kriging model fitting by Pentaspherical, the variogram of ANN fitting by Bessel model is better. The SSE of Kriging model is 0.0391, the SSE of ANN model is 0.0017. It shows the predicted map by ANN is more reliable.

## 12. Discussion

(3) The variogram is the central method of Geostatistics. It enables scientists to assess whether their data are spatially correlated and to what extent. With a suitable model for it. They can combine it with their data to predict by kriging, which in its simpler forms is one of weighted averaging. Kriging is an optimal method of prediction in that it provides unbiased estimates with minimum variance. However, it requires experiment to choose an appropriate model.

(4) Using the ANN gives the better fitting compare to original kriging (OK). But the distribution of ANN is not identical to OK. Therefore, the ideal that combines both ANN and OK to get the better spatial distribution. In the next steps, we will examine the cokriging-merge condition, and use it to compute the predicted map. It could outperform compare to separated method.

(5) In the article public in Korean Journal figured out the ANN is not suitable for spatial distribution compare to Kriging and IDW (Chun et al., 2019). We reviewed this paper because the comparison of ANN and Kriging for spatial analysis in South Korea relates to our study. The RMSE of OK is 6.662, and the RMSE of ANN is 14.607. It is contrasted to our results. Therefore, we propose to take the time to thorough analysis this paper and our results.

## References

- Chun, C., Choi, C., & Cho, J. (2019). 미시추 구간의 지반 층상정보 예측을 위한 정규 크리깅 및 인공신경망 기법의 비교. [Comparison of Ordinary Kriging and Artificial Neural Network for Estimation of Ground Profile Information in Unboring Region]. 20. doi:10.14481/JKGES.2019.20.3.15
- Pebesma, E. J. (2004). Multivariable geostatistics in S: the gstat package. *Computers & geosciences*, 30(7), 683-691. doi:<https://doi.org/10.1016/j.cageo.2004.03.012>
- Pebesma, E. J., & Wesseling, C. G. (1998). Gstat: a program for geostatistical modelling, prediction and simulation. *Computers & geosciences*, 24(1), 17-31. doi:[https://doi.org/10.1016/S0098-3004\(97\)00082-4](https://doi.org/10.1016/S0098-3004(97)00082-4)

## WEEK 26

A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network

### 13. Merging conditional model

Table 11. The status of processing data analysis

	Contents	Sub-contents
Week 1-2	<input checked="" type="checkbox"/> Data processing	<ul style="list-style-type: none"><li>21. Convert NDI to spatial-space format</li><li>22. Compute monthly average</li><li>23. Determine the specific time of the maximum value NDI</li><li>24. Retrieved precipitation from satellite</li><li>25. Retrieved temperature from satellite</li><li>26. Retrieved soil moisture from satellite</li><li>27. Retrieved elevation from satellite</li><li>28. Covert satellite-based data to grid-point and grid cell format</li><li>29. Analysis spatial trend of NDI</li></ul>
Week 3	<input type="checkbox"/> ANN model	<ul style="list-style-type: none"><li>30. Create surfaces for NDI, satellite-based data, longitude, latitude follow the spatial resolutions: 0.1 degree</li><li>31. Extract satellite-based precipitation, soil moisture, temperature, elevation, runoff, evapotranslation, longitude, latitude as 59 ASOS location. Check correlation of NDI and auxiliary variables</li><li>32. Set up the neural network regression (NNR) to predict NDI from auxiliary variables above</li><li>33. Build the structure of ANN</li><li>34. Training and validate ANN</li><li>35. Optimize the structure of ANN</li><li>36. Create a table to evaluate the model</li></ul>

	Contents	Sub-contents
Week 4	<input type="checkbox"/> Merging- conditional model	<p>37. Compute the residual 1 of ANN</p> <p>38. Compute residual 2 of Kriging regression</p> <p>39. Compute Cokring-conditional merging variogram of NN and Kriging</p> <p>40. Analysis spatial NDI coverage by using conditional variogram</p>

### 3.14. Compute Cokring-conditional merging variogram

In the previous steps, we determine the best model for kriging is a Pentaspherical model with sum square error value at 0.0391. Pentashpherical model follows the equation:

$$\begin{cases} \gamma(h) = \frac{15h}{a} - \frac{5}{4} \left(\frac{h}{a}\right)^3 + \frac{3}{8} \left(\frac{h}{a}\right)^5 \\ 0 \leq h \leq a \end{cases} \quad (1)$$

Where  $\gamma(h)$  is variogram function,  $h$  is range of data,  $a$  is the maximum range, that of the major direction of continuity (direction of spatial correlation at longest distances).

The best model kriging based on ANN is a Bessel model with sum square error value at 0.0017. Bessel model follows the equation:

$$\begin{cases} \gamma(h) = 1 - \frac{h}{a} K_1\left(\frac{h}{a}\right) \\ h \geq 0 \end{cases} \quad (2)$$

Where  $K_1$  is the first order modified Bessel functions. Using the variogram from two optimal models above, we computed maps of predictions. For combination variogram, we extract sample grid points from these maps.

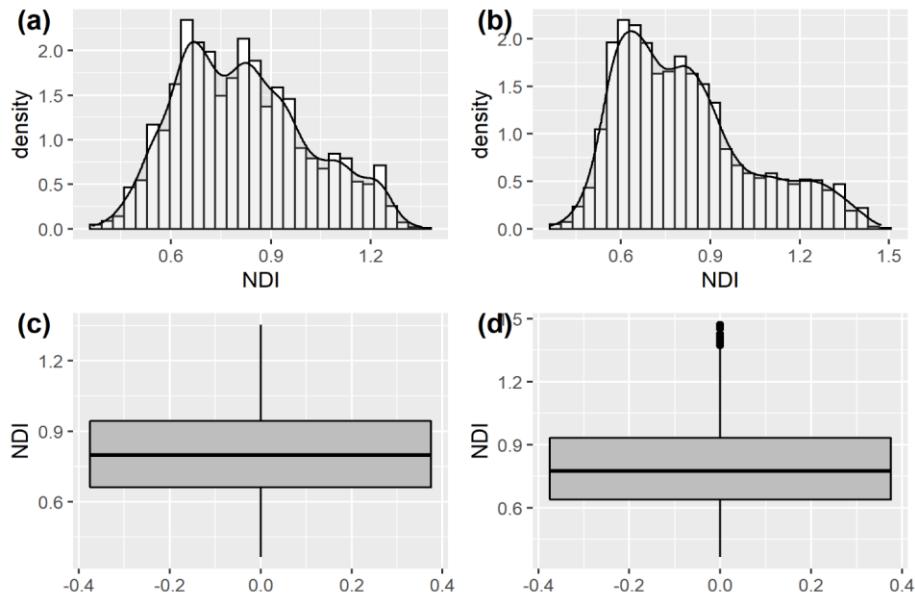


Fig 69. Estimate predicted NDI: (a) histogram and density of Kriging model, (b) histogram and density of ANN model, (c) boxplot of Kriging model, (d) boxplot of ANN model

Using the popular base distributions (Gamma, Lognormal, Empirical, log Empirical, student's t) fit of both data. Results show that the student's t distribution is the best fit for KR with the lowest AIC at -1709.337, mean value 0.815, standard deviation at 0.196. The ANN model matches to a Gamma distribution with the lowest AIC at -1172.916, mean value 0.817, standard deviation at 0.234.

The model covariance of variable from KR and ANN follow equation:

$$Y = 0.97162 * \text{Stable}(1.906, 2) \quad (3)$$

The nugget of model is approximate 0.00009, major range has value at 1.905, the lag size is 0.286.

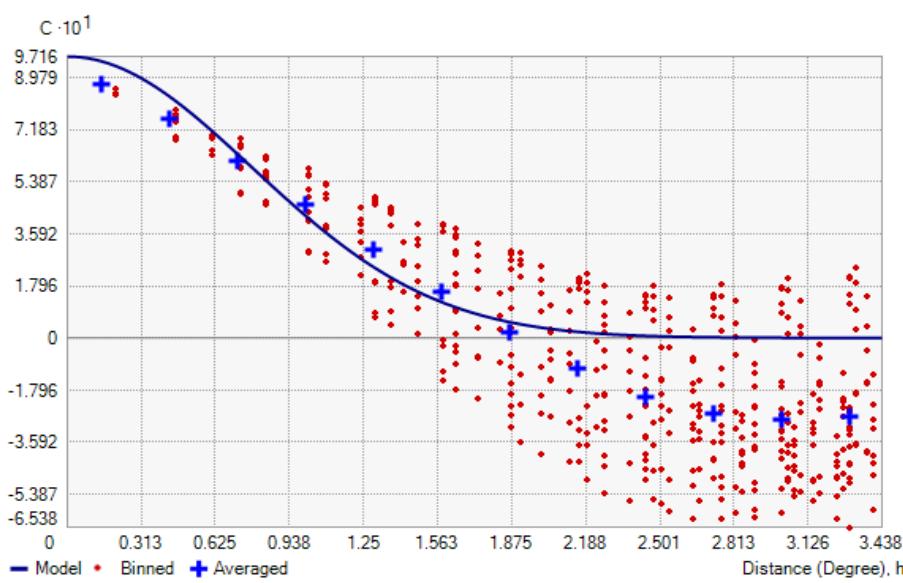


Fig 70. Covariance modelling of KR and ANN

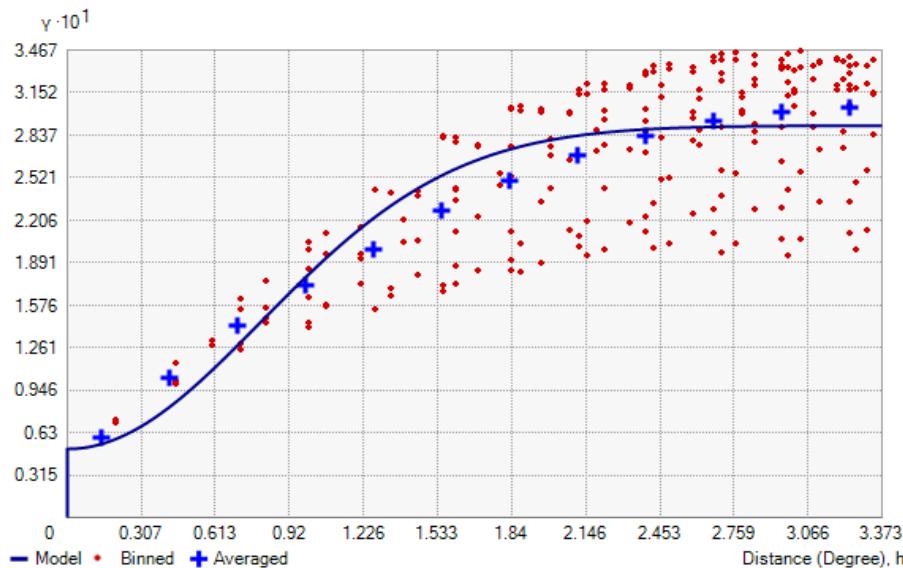
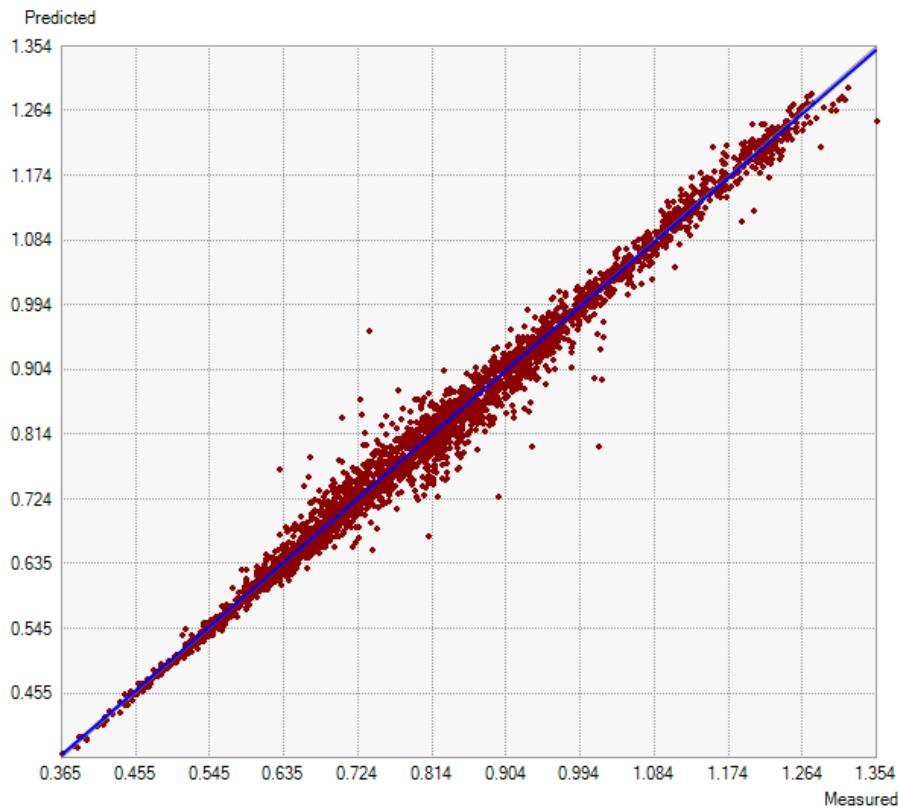


Fig 71. Semivariogram of KR and ANN model

Using KR model and ANN model, we create a regression model to predict NDI value. The regression model follows equation:

$$\text{NDI} = 0.990728285918355 * x + 0.00752529728149565 \quad (4)$$

Comparison of prediction and measurement is presented in Fig 4 below



*Fig 72. Estimate combine variogram of kriging and ANN*

The prediction errors include mean value at -0.000382, root mean square value at 0.0219, mean standardized value at 0.0395, root mean square standardized value at 3.148, and average standard value at 0.00674. This result was obtained from sample 3339 grid points. The grid point was extracted from kriging interpolation using Ordinary model and ANN model.

### 3.15. Analysis spatial NDI coverage by using conditional variogram

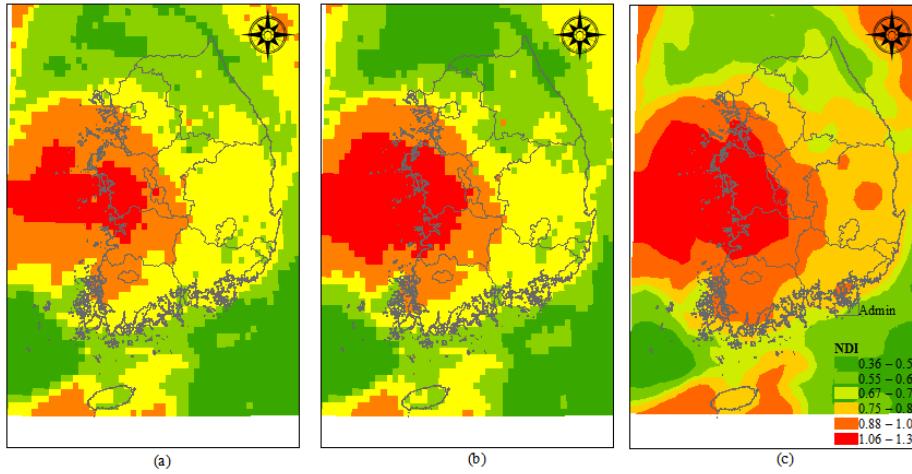


Fig 73. Spatial extreme drought coverage using: (a) original kriging, (b) ANN, (c) cokriging simple merge model

## 14. Discussion

(1) [Azimi, 2020 #4788] We were succeeded to propose the framework improve extreme drought coverage by using natural drought index, remote sensing data, and artificial neural network. The final predicted spatial drought map was obtained. The merging model based on the idea that keep the mean value at each model and adjust the variance by considering both residual of kriging model and residual of ANN. Simple combines variogram are examined by the fitting average value of both samples. This is the simplest method to combine variogram. However, the quantity of the accuracy of merging model compare to separate model is not clear. We propose to make the transparency of how much model is better. What is the suitable metric, indicator to present it? Beside that, we believe almost of the readers are interested in the reason why ANN could be used to improve the spatial distribution. Therefore, we propose to take the time to explain it.

(2) Another method could be considered in the future study that joint probability. We estimate the probability of each sample. Then use join probability function to combine these samples. In this case, we could use the copula functions to join distributions. However, this work could be repeated.

## WEEK 27

**A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network**

### 1. Find the metric to quantify the improvement of framework

We literature review the popular metric are used to evaluate a performance of model. Results are shown in the Table 1.

*Table 12. Some popular metrics are used to estimate spatial models*

No.	Titles	Authors	Metrics
01	Estimating surface soil moisture from satellite observations using a generalized regression neural network trained on sparse ground-based measurements in the continental U.S	(Yuan et al., 2020)	CV, bias, $R^2$ , RMSE, MAE
02	Estimating spatio-temporal air temperature in London (UK) using machine learning and earth observation satellite data	(dos Santos, 2020)	R, $R^2$ , MAE, RMSE
03	A hybrid kriging/land-use regression model with Asian culture-specific sources to assess NO <sub>2</sub> spatial-temporal variations	(Chen et al., 2020)	$R^2$ , adjust $R^2$
04	Comparison of spectral and spatial-based approaches for mapping the local variation of soil moisture in a semi-arid mountainous area	(Fathololoumi et al., 2020)	$R^2$ , RMSE
05	Artificial intelligence reconstructs missing climate information	(Kadow et al., 2020)	$R^2$ , RMSE
06	Digital mapping of soil carbon fractions with machine learning	(Keskin et al., 2019)	$R^2$ , RMSD, RI, RPD, RPIQ
07	A Method for the Optimized Design of a Rain Gauge Network Combined with Satellite Remote Sensing Data	(Huang et al., 2020)	ISA
08	Empirical Bayesian kriging implementation and usage	(Gribov & Krivoruchko, 2020)	RMSPE
09	Examination of geostatistical and machine-learning techniques as interpolators in anisotropic atmospheric environments	(Tadić et al., 2015)	CV, RRMSE, RMAE
10	Modeling Short Term Rainfall Forecast Using Neural Networks, and Gaussian Process Classification Based on the SPI Drought Index	(Azimi & Moghaddam, 2020)	ME, RMSE,

No.	Titles	Authors	Metrics
			SME, SRMSE

In table 1, CV presents for Cross validate.  $R^2$  is the correlation coefficient. RMSE is the root mean square error. MAE is the mean absolute error. R is the Pearson correlation. RMSD is the root mean square deviation. RI is the relative decrease in RMSD. RPD is the Residual prediction deviation. RPIQ is the ratio of prediction error to inter-quartile range. ISA is the information content, spatiotemporally, and accuracy. RMSPE is the root-mean-squared prediction error. RMAE is the relative mean absolute error, RRMSE is the relative root mean square error. ME is the mean error. SME is the standardized mean error. SRMSE is the standardized root mean square error.

We found that CV, RMSE, MAE are the most popular metrics. CV is used to ensure a model is robust. A portion of the data is held back, while the bulk of data is trained, and holdout sample is used to test the model. CV using multiple tests by sequential holdout samples that cover all of the data. For instance, in k-fold cross validation, the data is divided into k random subsets. A total of k models is fitted, and k validation statistics are obtained. The CV can use on the datasets ranging from small to large size. In the small data set, CV is efficient of exploited the limited data (Dangeti, 2017). CV could be used on models where iterative processes responding to local structures (Bruce et al., 2020).

The  $R^2$  is correlation coefficient using commonly in the linear regression. It is used to find how a strong relationship between the data. The range of  $R^2$  from 0 to 1. The 0 shows the non-correlation. The value  $R^2$  at 1 show the absolute correlation.

RMSE is the standard deviation of residuals. It shows the distance from the regression line to the observational points. In another word, it shows the concentration of data around the best fit (Kenney, 1939).

MAE is the average of all absolute error. After comparing RMSE and MAE, Willmott and Matsuura (2005) indicate that MAE is a more natural measure of average error. Dimensioned evaluations and inter-comparisons of average model-performance error, therefore, should be based on MAE.

We decided using CV,  $R^2$ , RMSE, MAE as these metrics for our study. 10 kfolds of CV was used to estimate model using root square correlation (RSQ), RMSE, MAE. Results show that the maximum of RMSE is at 0.26, the maximum of MAE is at 0.015 and the lowest correlation at 0.982 (Table 2). These results indicate cokriging of KR and ANN can be used to estimate extreme drought coverages.

Table 13. Model evaluation

KFOLD	RMSE	MAE	RSQ
1	0.020	0.014	0.989
2	0.022	0.015	0.987
3	0.023	0.015	0.986
4	0.022	0.015	0.987
5	0.024	0.015	0.984
6	0.026	0.015	0.982
7	0.020	0.014	0.991
8	0.018	0.013	0.991
9	0.020	0.013	0.989
10	0.023	0.015	0.987

## 2. Explain reasons our framework can improve spatial drought coverage

In this study, we figured out the reason of ANN can improve the spatial drought coverages. The ANN is outperformed to the traditional methods. Because it can learn from non-linearity characteristic. Adapting weight, active functions, connecting the neurons help model learn better. The correlation of final model is higher than the using the linear correlation of data.

Furthermore, the spatial characteristic was restrained by residual kriging model. It keeps the mean of a sample and minimize the deviation is the key idea for spatial prediction. Combining ANN and Kriging residues in cokriging model is achieved both spatial and non-linear characteristic.

## 3. Discussion

(1) The correlation of the model is very high. We should double check the computation of it.

Beside that we should be aware of correlation showing the relationship of observation and prediction. It is hard to evaluate model based only on correlation. Therefore, combine several metrics is the better model for assessment.

(2) Although, the results of model are not perfect at the present, we are succeeding in providing the framework to estimate spatial extreme drought coverage. We proposed a framework using natural drought index, satellite-based data and artificial neural network. The uniqueness of our study that combine a statistic spatial model and neural network to estimate spatial extreme drought

that have lacked attentions before. In the next step, we propose writing two pages summary of our study before writing an article.

## References

- Azimi, S., & Moghaddam, M. A. (2020). Modeling Short Term Rainfall Forecast Using Neural Networks, and Gaussian Process Classification Based on the SPI Drought Index. *Water resources management*, 34(4), 1369-1405. doi:10.1007/s11269-020-02507-6
- Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*: O'Reilly Media.
- Chen, T.-H., Hsu, Y.-C., Zeng, Y.-T., Candice Lung, S.-C., Su, H.-J., Chao, H. J., & Wu, C.-D. (2020). A hybrid kriging/land-use regression model with Asian culture-specific sources to assess NO<sub>2</sub> spatial-temporal variations. *Environmental Pollution*, 259, 113875. doi:<https://doi.org/10.1016/j.envpol.2019.113875>
- Dangeti, P. (2017). *Statistics for machine learning*: Packt Publishing Ltd.
- dos Santos, R. S. (2020). Estimating spatio-temporal air temperature in London (UK) using machine learning and earth observation satellite data. *International Journal of Applied Earth Observation and Geoinformation*, 88, 102066. doi:<https://doi.org/10.1016/j.jag.2020.102066>
- Fathololoumi, S., Vaezi, A. R., Alavipanah, S. K., Ghorbani, A., & Biswas, A. (2020). Comparison of spectral and spatial-based approaches for mapping the local variation of soil moisture in a semi-arid mountainous area. *Science of The Total Environment*, 724, 138319. doi:<https://doi.org/10.1016/j.scitotenv.2020.138319>
- Gribov, A., & Krivoruchko, K. (2020). Empirical Bayesian kriging implementation and usage. *Science of The Total Environment*, 722, 137290. doi:<https://doi.org/10.1016/j.scitotenv.2020.137290>
- Huang, Y., Zhao, H., Jiang, Y., & Lu, X. (2020). A Method for the Optimized Design of a Rain Gauge Network Combined with Satellite Remote Sensing Data. *Remote Sensing*, 12(1), 194.
- Kadow, C., Hall, D. M., & Ulbrich, U. (2020). Artificial intelligence reconstructs missing climate information. *Nature Geoscience*. doi:10.1038/s41561-020-0582-5
- Kenney, J. F. (1939). *Mathematics of statistics*: D. Van Nostrand.
- Keskin, H., Grunwald, S., & Harris, W. G. (2019). Digital mapping of soil carbon fractions with machine learning. *Geoderma*, 339, 40-58. doi:<https://doi.org/10.1016/j.geoderma.2018.12.037>
- Tadić, J. M., Ilić, V., & Biraud, S. (2015). Examination of geostatistical and machine-learning techniques as interpolators in anisotropic atmospheric environments. *Atmospheric Environment*, 111, 28-38. doi:<https://doi.org/10.1016/j.atmosenv.2015.03.063>
- Willmott, C. J., & Matsuura, K. J. C. r. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. 30(1), 79-82.

Yuan, Q., Xu, H., Li, T., Shen, H., & Zhang, L. (2020). Estimating surface soil moisture from satellite observations using a generalized regression neural network trained on sparse ground-based measurements in the continental U.S. *Journal of Hydrology*, 580, 124351.  
doi:<https://doi.org/10.1016/j.jhydrol.2019.124351>

## WEEK 28

### A framework evaluates spatial extreme drought coverage using natural drought index, Satellite based data and artificial neural network

#### Motivations:

- Drought is one of the most severe natural disasters. Different with others, it is a creeping phenomenon and occurring in huge region. Spatial extreme drought analysis is vital to understand the characteristics of regional drought, and to provide a reference for drought damage reduction. Several studies have committed to figure out spatial characteristic of extreme drought.
- The spatial characteristic of drought is usually interpolated from gauge stations data. The comprehensive comparison of various interpolation was presented in studies (Li and Heap, 2011; Meng et al., 2013). IDW and OK are the most frequently used methods. OK was used to analyze spatial extreme drought for China from 1961-2013 with 810 weather stations (Liu et al., 2016). IDW was used to derive spatial extreme drought for South Korea from 1981-2016 with 59 weather stations (Vo et al., 2020). The spatial extreme drought analysis based on interpolated methods are fast, simple algorithms. However, the accuracy of interpolation depends on the weather stations density that is difficult to be obtained in the reality.
- Intensive observed data obtained from other sources are also available as the auxiliary variables. To utilize the auxiliary information in these data, methods such as regression kriging (RK) or cokriging are proposed. But these methods all assume that the auxiliary variables keep linear correlation with the target variable implicitly, which is not satisfied in most cases. To improve the interpolation for the region where has limited gauges data, the satellite data is usually utilized to surplus to gauges data. Although of poor quality in value compare to gauges data, the remote sensing data is better on spatial distribution. The combination of satellite-based data and gauge-based data can improve regional drought monitoring and assessment (Bai et al., 2019).
- Artificial Neural network (ANN) is a method using computer to model human brain leaning. ANN have been used in spatial analysis. Extracting the feature of remote sensing data for merging with gauge-based data, machine learning was utilized. Because it can obtain the complexity spatial pattern of spatial characteristics. Recently, several studies using machine learning approach, satellite data to improve spatial distribution such as digital mapping of soil carbon fraction (Keskin et al., 2019), air temperature (dos Santos, 2020). Regression kriging combine neural network kriging (RKNNRK) is predict spatial air temperature (Yuan et al., 2020).
- Previous studies above are success showing the capacity of Kriging and ANN to develop the spatial distribution of air temperature, soil moisture. However, application of spatial extreme drought coverages has been lacked attention.

#### Objectives:

- The aim of this study is development a framework for spatial extreme drought assessment. It is used to improve spatial extreme drought distribution by sing natural drought index, satellite-based data, and ANN. The framework remains the mean value of original data and try to minimize the deviation.

#### Methodology:

- The methodology for spatial extreme drought coverage is summarized in **Fig 1**. The procedure consists of 5 steps. Step 1 and 2 build VIC model to extract runoff, soil moisture. Step 3 computes NDI. Step 4 setup the coupling Artificial Neural Network and Kriging model to predict spatial drought coverage, and final step 5 estimate results by using cross validation of correlation, root mean square error, absolute mean error.
- The input data of VIC model are soil parameters, vegetation parameters, vegetation library, meteorological forcing. The digital elevation model (DEM) was provided by Ministry of Land. The meteorological data included 59 ASOS. Total 586 grid cell at 1/80 (12.5 km) was generated

for VIC model. Maximum, minimum temperature and average wind speed was collected at each weather stations. They are interpolated to build the grid with the same spatial 12.5 km (Son et al., 2011).

- VIC model was calibrated and validated by comparing the stream flow at dam operation with observational discharge from Water Resource Management Information System (WMIS). These outputs of VIC model were accumulated to monthly scale as these input data for computed NDI.
- NDI was computed by using principle component analysis (PCA). Precipitation, Runoff, Soil moisture are the input of NDI computation. We computed NDI at 59 ASOS. ASOS provide the precipitation. The runoff and soil moisture are extracted from grid cells. These grid cells are matched to ASOS by closest neighbor method. PCA is used to reduce dimension methods. It based on the linearity that determine the new vector that contains the most information from input vectors.
- Multiple regression using ANN to create the first drought map. The second drought map is computed by simple kriging. Then the coupling two spatial drought map by cokriging model. Finally, the coupling model is evaluated by cross validation.

#### **Study area and data:**

- Study area: Study area is South Korea with a total area of 100,032 square km. Data: Precipitation data was collected from 59 Automatic Synoptic Observation System (ASOS) of Korea Meteorological Agency. Monthly rainfall data are monitored from 1981 to 2016. **Fig 2** presents Study area and location of ASOS. Runoff and Soil moisture data were also simulated for the same period. Satellite-based data: precipitation (p), runoff (r), soil moisture (s), evapotranspiration (e), temperature (t), win speed (w), topographical elevation (z).

#### **Results:**

- The most suitable number of hidden neurons (5), active function (swish), and auxiliary variable (p, r, s, e, t) using for optimal ANN was presented in **Fig3**.
- **Table 1** presents the best fit for variogram of Kriging model and ANN model. Based on the sum square error (SSE), The Pentaspherical is the best fit model for Kriging (0.0391). While the Bessel model is the best fit for ANN (0.0017).
- Using the popular base distributions (Gamma, Lognormal, Empirical, log Empirical, student's t) fit of both data. Results show that the student's t distribution is the best fit for KR with the lowest AIC at -1709.337, mean value 0.815, standard deviation at 0.196. The ANN model matches to a Gamma distribution with the lowest AIC at -1172.916, mean value 0.817, standard deviation at 0.234.
- The semi variogram and coupling variogram of KR and ANN was present in the **Fig 4**.
- Model evaluation based on cross validation, root mean square error, mean absolute error, root square is presented in **Table 2**.
- The spatial extreme drought coverage using kriging, ANN and cokriging model were presented in **Fig 5**.

#### **Conclusions:**

- Determining the spatial extreme drought coverage is one of the most important issues in water resource planning and management.
- This study analyzed spatial extreme drought coverage using natural drought index, satellite-based data, and artificial neural network.
- The study has analyzed the correlation of spatial drought and hydrometeorological data such as precipitation, runoff, soil moisture, evapotranspiration, temperature, win speed, topographical elevation. These data could be obtained by using remote-sensing sources.
- Although, this study used the simplest artificial intelligence-ANN, results show combination of the geo statistic model kriging and ANN is outperform compare to a separate model. Therefore,

hybrid model of geo-statistic and deep learning could be a promised method to improve spatial extreme drought coverage.

## Appendix

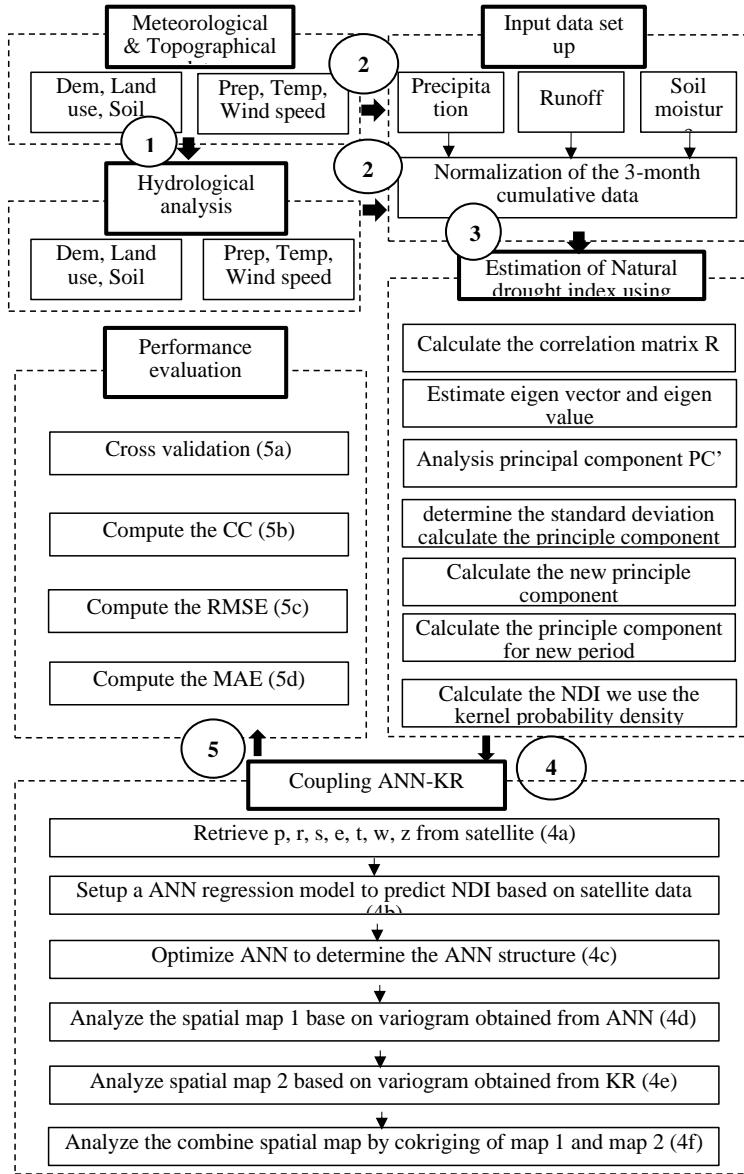


Fig 74. Framework of spatial extreme drought coverage assessment

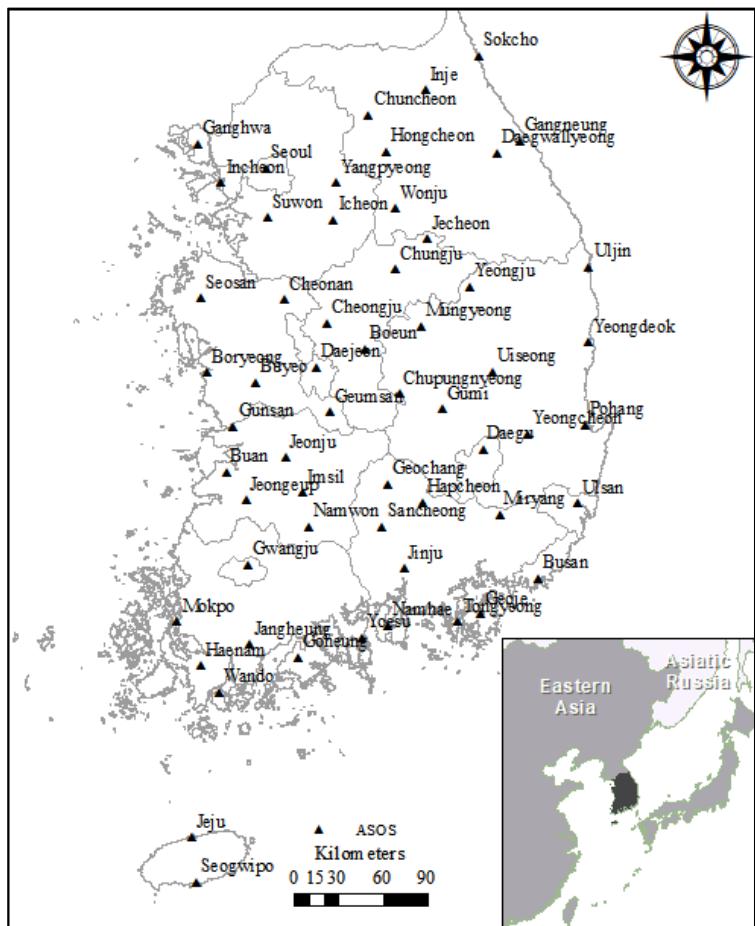


Fig 75. ASOS locations and study area

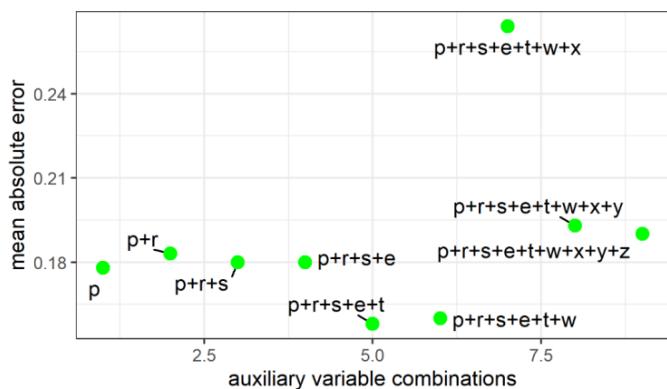
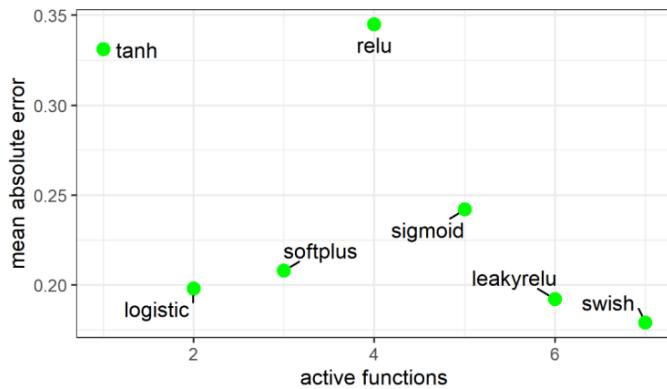
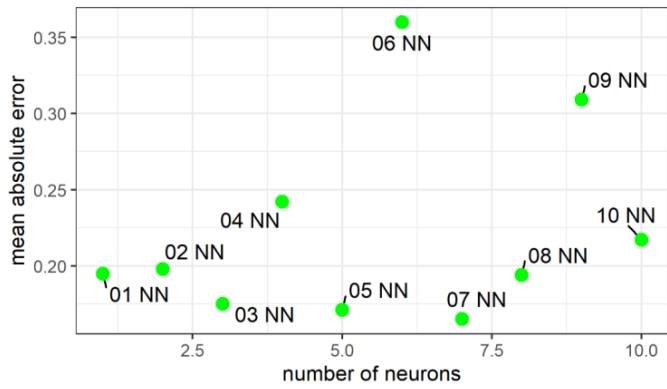
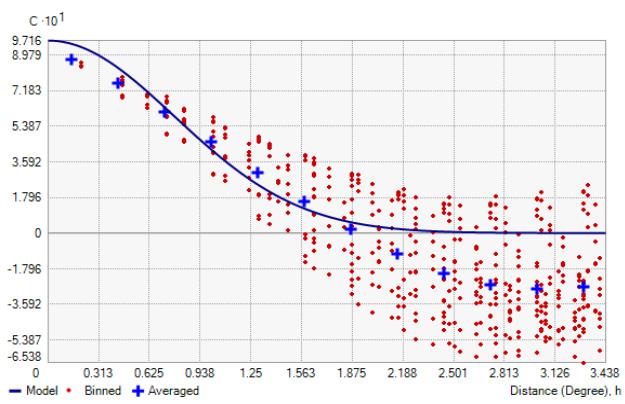


Fig 76. Optimize ANN structures with number of neurons, active functions, and auxiliary variables



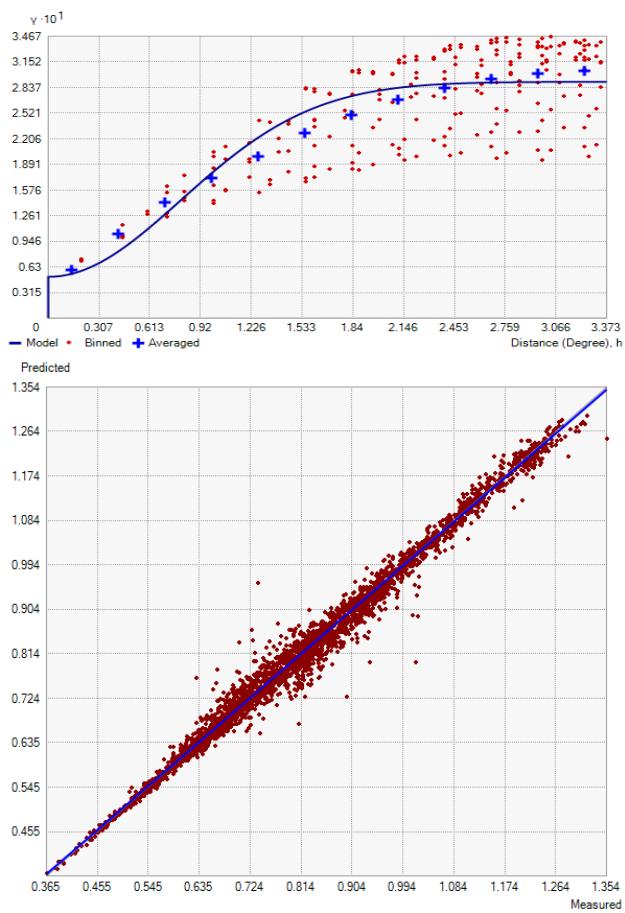


Fig 77. Covariance modeling, semi variogram, and scatter diagram of kriging and ANN model

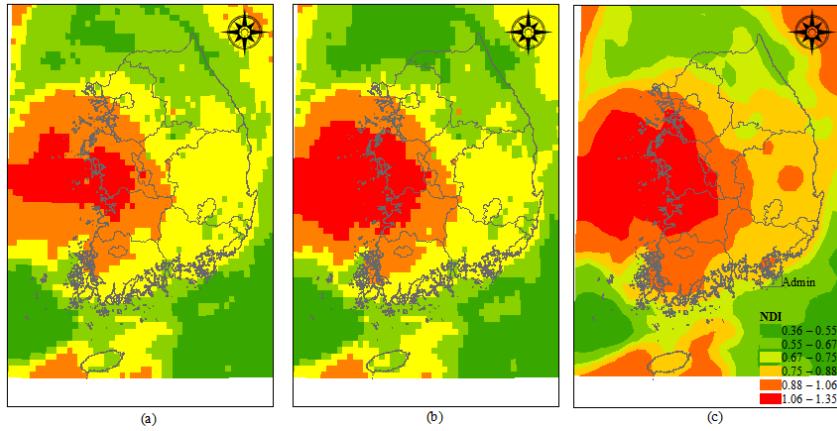


Fig 78. Spatial extreme drought coverage using: (a) original kriging, (b) ANN, (c) cokriging simple merge model

Table 14. Determine the variogram model for ANN and kriging

ANN model			Kriging model	
No.	Model	Sum square error	Model	Sum square error
1	Bessel	0.0017	Pentaspherical	0.0391
2	Exponential	0.0021	Spherical	0.0408
3	Gaussian	0.0095	Linear	0.0421
4	Linear	0.0019	Exponential	0.0478
5	Pentaspherical	0.0033	Gaussian	0.0507
6	Spherical	0.0028	Bessel	0.0607

Table 15. Model evaluation

KFOLD	RMSE	MAE	RSQ
1	0.020	0.014	0.989
2	0.022	0.015	0.987
3	0.023	0.015	0.986
4	0.022	0.015	0.987
5	0.024	0.015	0.984
6	0.026	0.015	0.982
7	0.020	0.014	0.991
8	0.018	0.013	0.991
9	0.020	0.013	0.989
10	0.023	0.015	0.987

**Professor comments: make clear methododology**

**NDI 01** là giá trị tính toán tại 59 trạm. Giá trị **NDI 02** là giá trị nội suy từ 59 trạm theo IDW.

-Q1: NDI nào được sử dụng để training mô hình?

-Q2: Dữ liệu đầu nào để xây dựng bản đồ NDI 3 (bản đồ 1)?

-Q3: Dữ liệu nào để xây dựng bản đồ NDI 4 (bản đồ 2)?

-Q4: làm cách nào để biết được phương pháp này hiệu quả hơn?

**Brain storm**

-Q1: NDI 1 được dùng để training mô hình.

-Q2: NDI3 đạt được từ regression từ ANN ra các vị trí không có giá trị. Trainning NDI 1 tạo ra các giá trị khác mà NDI 01 không có.

-Q3: từ NDI 01 sử dụng SK để nội suy ra giá trị NDI 3 (bản đồ 2). Dùng cokriging để tạo ra NDI 4 .

-Q4: Để đánh giá mô hình ta so sánh giá trị NDI 1 (59 điểm) với giá trị tự mô hình dự báo tại 59 điểm này. Vì các điểm khác nhau không có liệu để so sánh.

## WEEK 29

### A framework evaluates spatial extreme drought coverage using natural drought index, Satellite based data and artificial neural network

#### Motivations:

- Drought is one of the most severe natural disasters. Different with others, it is a creeping phenomenon and occurring in huge region. Spatial extreme drought analysis is vital to understand the characteristics of regional drought, and to provide a reference for drought damage reduction. Several studies have committed to figure out spatial characteristic of extreme drought.
- The spatial characteristic of drought is usually interpolated from gauge stations data. The comprehensive comparison of various interpolation was presented in studies (Li and Heap, 2011; Meng et al., 2013). IDW and OK are the most frequently used methods. OK was used to analyze spatial extreme drought for China from 1961-2013 with 810 weather stations (Liu et al., 2016). IDW was used to derive spatial extreme drought for South Korea from 1981-2016 with 59 weather stations (Vo et al., 2020). The spatial extreme drought analysis based on interpolated methods are fast, simple algorithms. However, the accuracy of interpolation depends on the weather stations density that is difficult to be obtained in the reality.
- Intensive observed data obtained from other sources are also available as the auxiliary variables. To utilize the auxiliary information in these data, methods such as regression kriging (RK) or cokriging are proposed. But these methods all assume that the auxiliary variables keep linear correlation with the target variable implicitly, which is not satisfied in most cases. To improve the interpolation for the region where has limited gauges data, the satellite data is usually utilized to surplus to gauges data. Although of poor quality in value compare to gauges data, the remote sensing data is better on spatial distribution. The combination of satellite-based data and gauge-based data can improve regional drought monitoring and assessment (Bai et al., 2019).
- Artificial Neural network (ANN) is a method using computer to model human brain leaning. ANN have been used in spatial analysis. Extracting the feature of remote sensing data for merging with gauge-based data, machine learning was utilized. Because it can obtain the complexity spatial pattern of spatial characteristics. Recently, several studies using machine learning approach, satellite data to improve spatial distribution such as digital mapping of soil carbon fraction (Keskin et al., 2019), air temperature (dos Santos, 2020). Regression kriging combine neural network kriging (RKNRK) is predict spatial air temperature (Yuan et al., 2020).
- Previous studies above are success showing the capacity of Kriging and ANN to develop the spatial distribution of air temperature, soil moisture. However, application of spatial extreme drought coverages has been lacked attention.

#### Objectives:

- The aim of this study is development a framework for spatial extreme drought assessment. It is used to improve spatial extreme drought distribution by sing natural drought index, satellite-based data, and ANN. The framework remains the mean value of original data and try to minimize the deviation.

#### Methodology:

- The methodology for spatial extreme drought coverage is summarized in **Fig 1**. The procedure consists of 5 steps. Step 1 and 2 build VIC model to extract runoff, soil moisture. Step 3 computes NDI. Step 4 setup the coupling Artificial Neural Network and Kriging model to predict spatial drought coverage, and final step 5 estimate results by using cross validation of correlation, root mean square error, absolute mean error.
- The input data of VIC model are soil parameters, vegetation parameters, vegetation library, meteorological forcing. The digital elevation model (DEM) was provided by Ministry of Land.

The meteorological data included 59 ASOS. Total 586 grid cell at 1/80 (12.5 km) was generated for VIC model. Maximum, minimum temperature and average wind speed was collected at each weather stations. They are interpolated to build the grid with the same spatial 12.5 km (Son et al., 2011).

- VIC model was calibrated and validated by comparing the stream flow at dam operation with observational discharge from Water Resource Management Information System (WMIS). These outputs of VIC model were accumulated to monthly scale as these input data for computed NDI.
- NDI was computed by using principle component analysis (PCA). Precipitation, Runoff, Soil moisture are the input of NDI computation. We computed NDI at 59 ASOS. ASOS provide the precipitation. The runoff and soil moisture are extracted from grid cells. These grid cells are matched to ASOS by closest neighbor method. PCA is used to reduce dimension methods. It based on the linearity that determine the new vector that contains the most information from input vectors.
- Multiple regression using ANN to create the first drought map. The second drought map is computed by simple kriging. Then the coupling two spatial drought map by cokriging model. Finally, the coupling model is evaluated by cross validation.

#### **Study area and data:**

- Study area: Study area is South Korea with a total area of 100,032 square km. Data: Precipitation data was collected from 59 Automatic Synoptic Observation System (ASOS) of Korea Meteorological Agency. Monthly rainfall data are monitored from 1981 to 2016. **Fig 2** presents Study area and location of ASOS. Runoff and Soil moisture data were also simulated for the same period. Satellite-based data: precipitation (p), runoff (r), soil moisture (s), evapotranspiration (e), temperature (t), win speed (w), topographical elevation (z).

#### **Results:**

- The most suitable number of hidden neurons (5), active function (swish), and auxiliary variable (p, r, s, e, t) using for optimal ANN was presented in **Fig3**.
- **Table 1** presents the best fit for variogram of Kriging model and ANN model. Based on the sum square error (SSE), The Pentaspherical is the best fit model for Kriging (0.0391). While the Bessel model is the best fit for ANN (0.0017).
- Using the popular base distributions (Gamma, Lognormal, Empirical, log Empirical, student's t) fit of both data. Results show that the student's t distribution is the best fit for KR with the lowest AIC at -1709.337, mean value 0.815, standard deviation at 0.196. The ANN model matches to a Gamma distribution with the lowest AIC at -1172.916, mean value 0.817, standard deviation at 0.234.
- The semi variogram and coupling variogram of KR and ANN was present in the **Fig 4**.
- Model evaluation based on cross validation, root mean square error, mean absolute error, root square is presented in **Table 2**.
- The spatial extreme drought coverage using kriging, ANN and cokriging model were presented in **Fig 5**.

#### **Conclusions:**

- Determining the spatial extreme drought coverage is one of the most important issues in water resource planning and management.
- This study analyzed spatial extreme drought coverage using natural drought index, satellite-based data, and artificial neural network.
- The study has analyzed the correlation of spatial drought and hydrometeorological data such as precipitation, runoff, soil moisture, evapotranspiration, temperature, win speed, topographical elevation. These data could be obtained by using remote-sensing sources.
- Although, this study used the simplest artificial intelligence-ANN, results show combination of the geo statistic model kriging and ANN is outperform compare to a separate model. Therefore,

hybrid model of geo-statistic and deep learning could be a promised method to improve spatial extreme drought coverage.

## Appendix

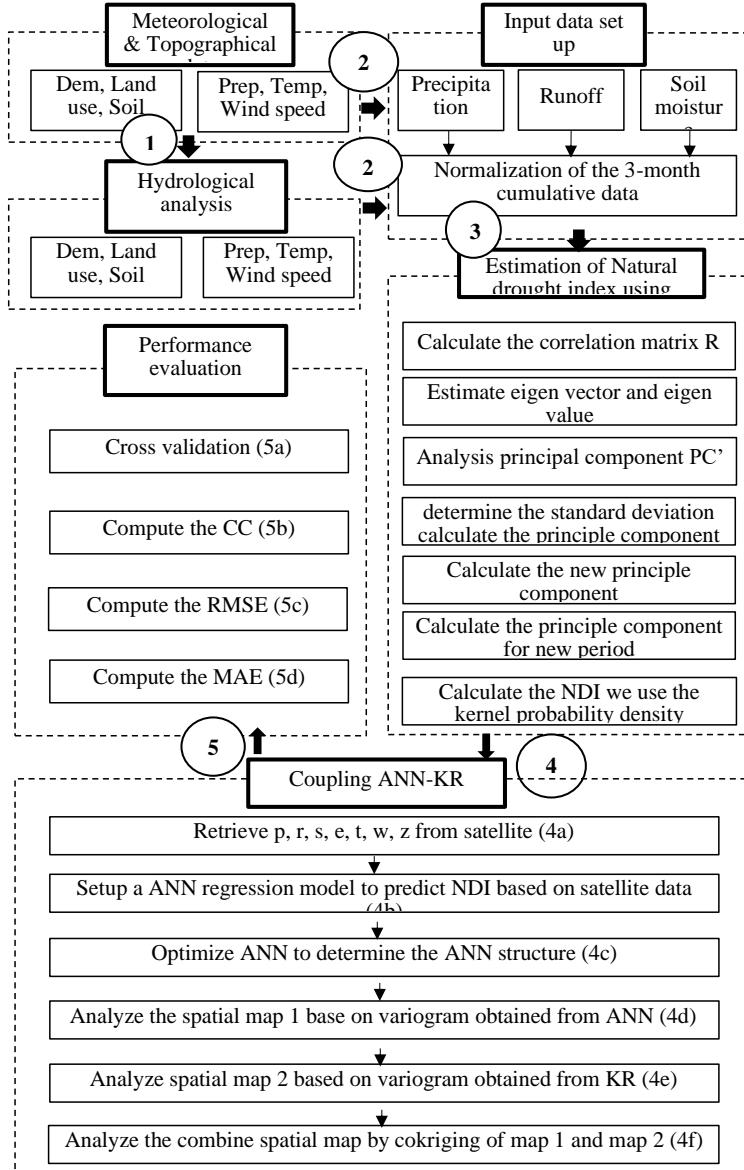


Fig 79. Framework of spatial extreme drought coverage assessment

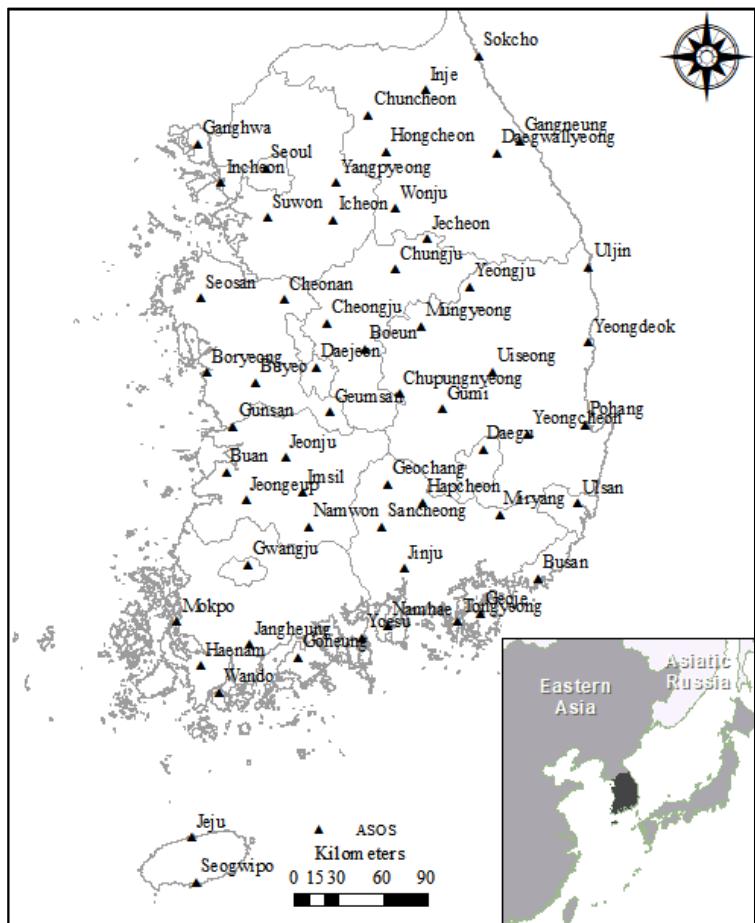


Fig 80. ASOS locations and study area

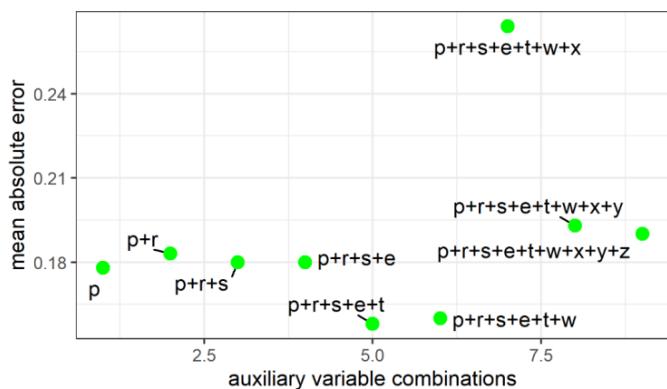
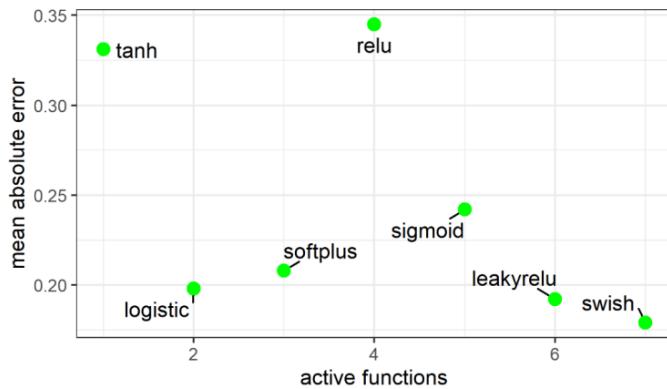
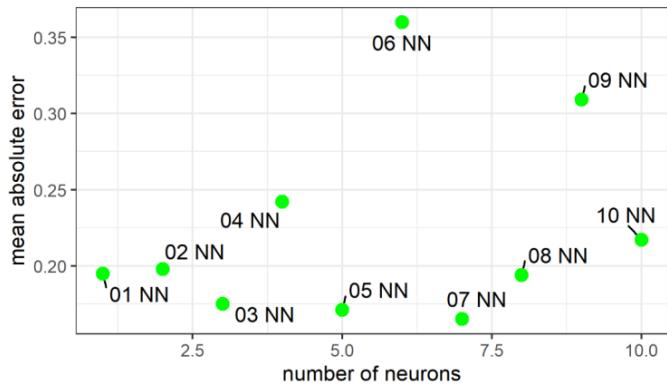
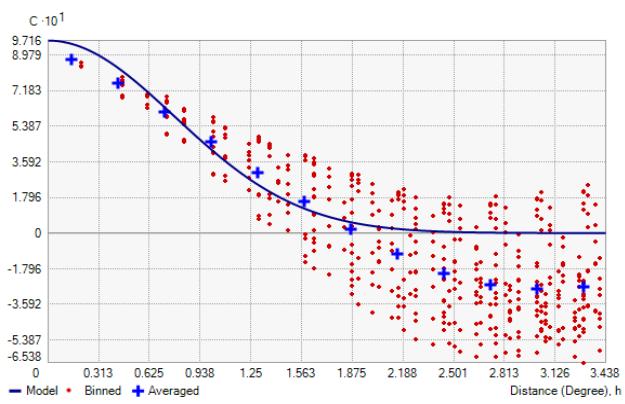


Fig 81. Optimize ANN structures with number of neurons, active functions, and auxiliary variables



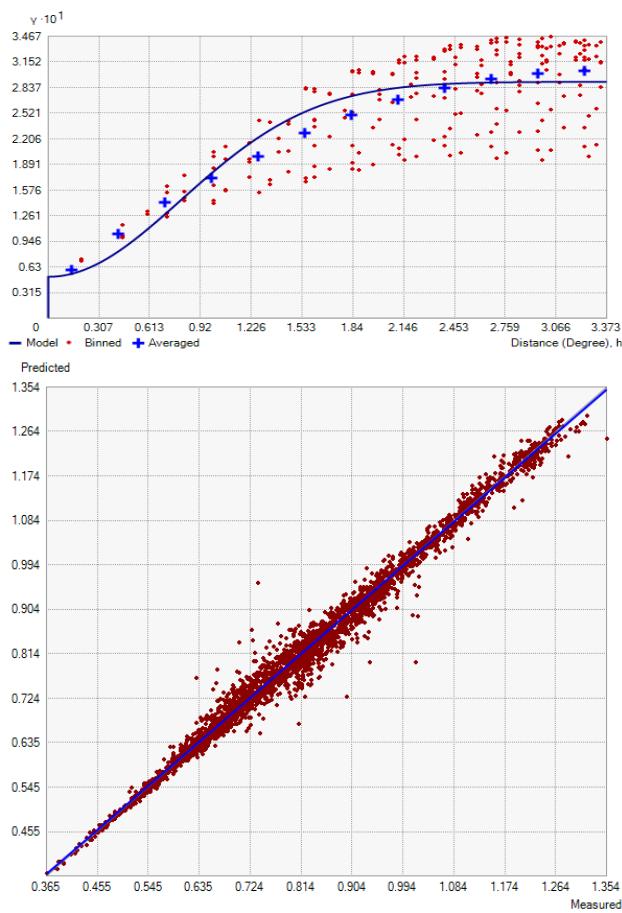


Fig 82. Covariance modeling, semi variogram, and scatter diagram of kriging and ANN model

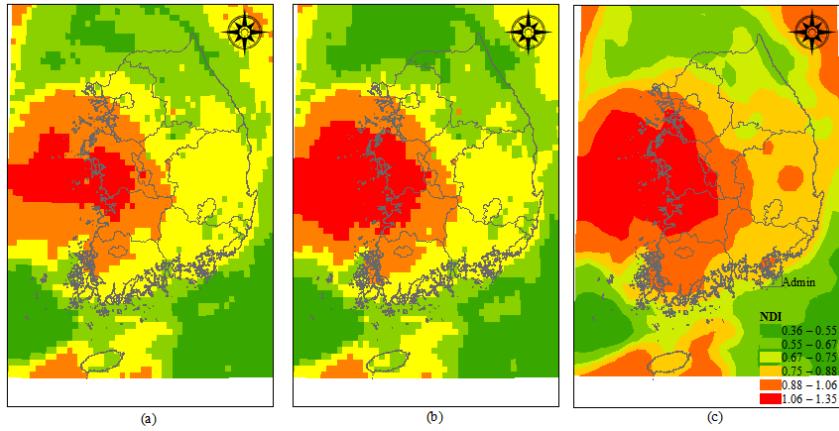


Fig 83. Spatial extreme drought coverage using: (a) original kriging, (b) ANN, (c) cokriging simple merge model

Table 16. Determine the variogram model for ANN and kriging

ANN model			Kriging model	
No.	Model	Sum square error	Model	Sum square error
1	Bessel	0.0017	Pentaspherical	0.0391
2	Exponential	0.0021	Spherical	0.0408
3	Gaussian	0.0095	Linear	0.0421
4	Linear	0.0019	Exponential	0.0478
5	Pentaspherical	0.0033	Gaussian	0.0507
6	Spherical	0.0028	Bessel	0.0607

*Table 17. Model evaluation*

<b>KFOLD</b>	<b>RMSE</b>	<b>MAE</b>	<b>RSQ</b>
1	0.020	0.014	0.989
2	0.022	0.015	0.987
3	0.023	0.015	0.986
4	0.022	0.015	0.987
5	0.024	0.015	0.984
6	0.026	0.015	0.982
7	0.020	0.014	0.991
8	0.018	0.013	0.991
9	0.020	0.013	0.989
10	0.023	0.015	0.987

## WEEK 30

A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network

### 4. Clarify these unclear parts of the methodology

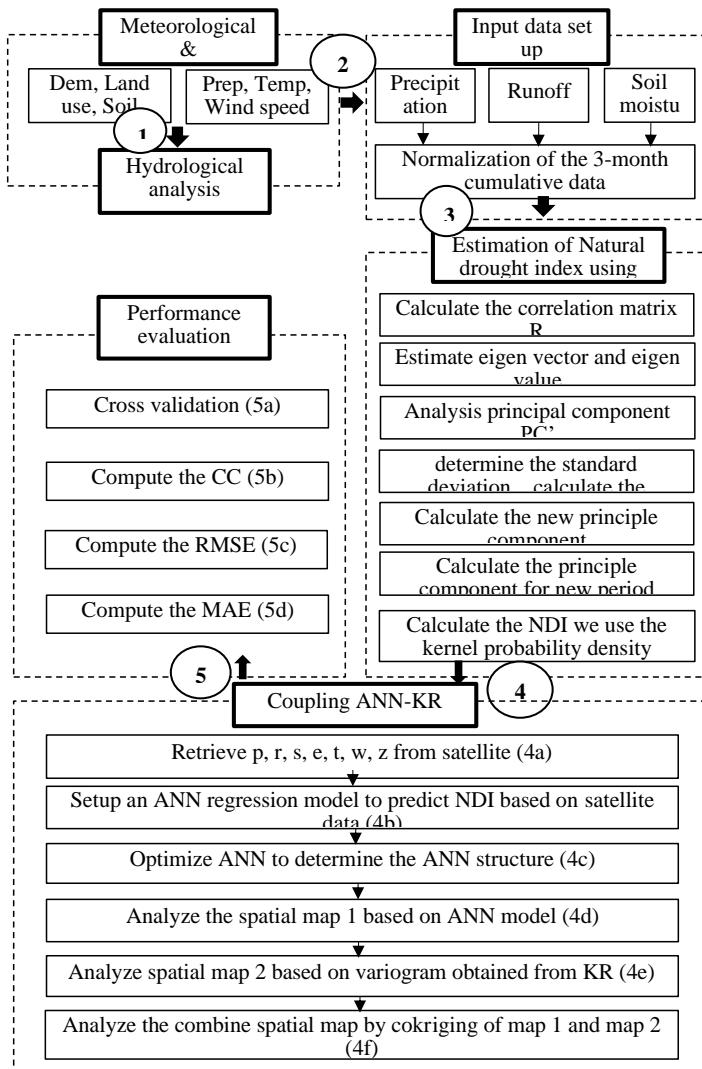
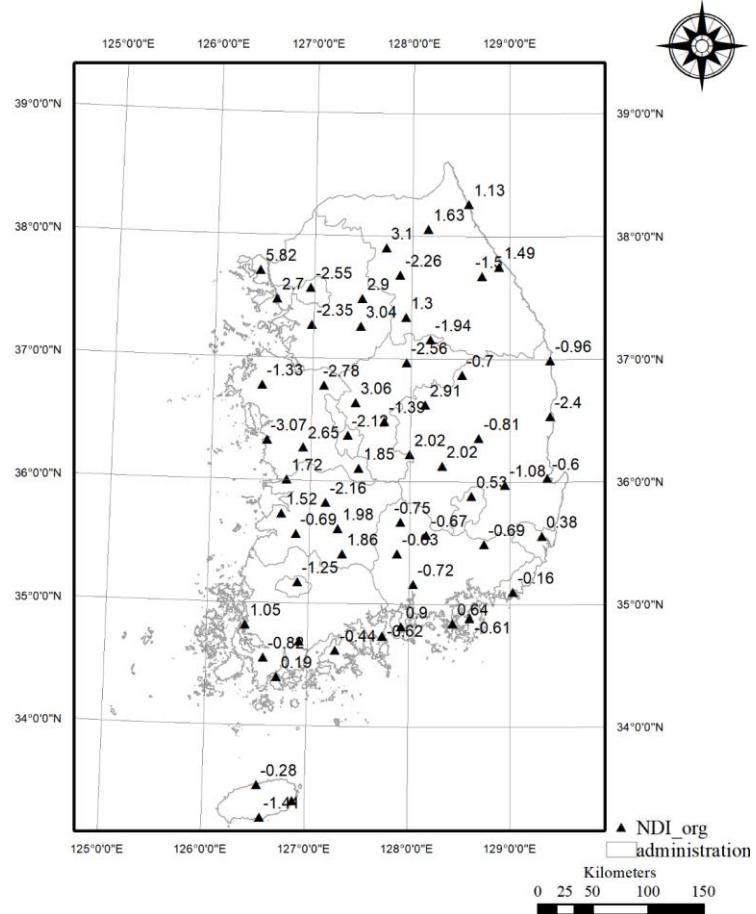


Fig 84.

Complete step 3, we got the extreme drought at 59 ASOS stations in August 2015 (Fig 2).



*Fig 85. The original NDI computed from precipitation, runoff, soil moisture.*

In the previous study, we used the invert distance weight (IDW) to interpolate the spatial drought map. Results from IDW depend on the distance of original NDI points and predicted NDI locations. The predicted points are closer the original points are better estimation. While the other predicted points that are far from original points are difficult to predict accurately. As the name, map from IDW has a trend to create “islands” with centroid are the original points (Fig 3). Furthermore, IDW does not consider the spatial autocorrelation of predicted map.

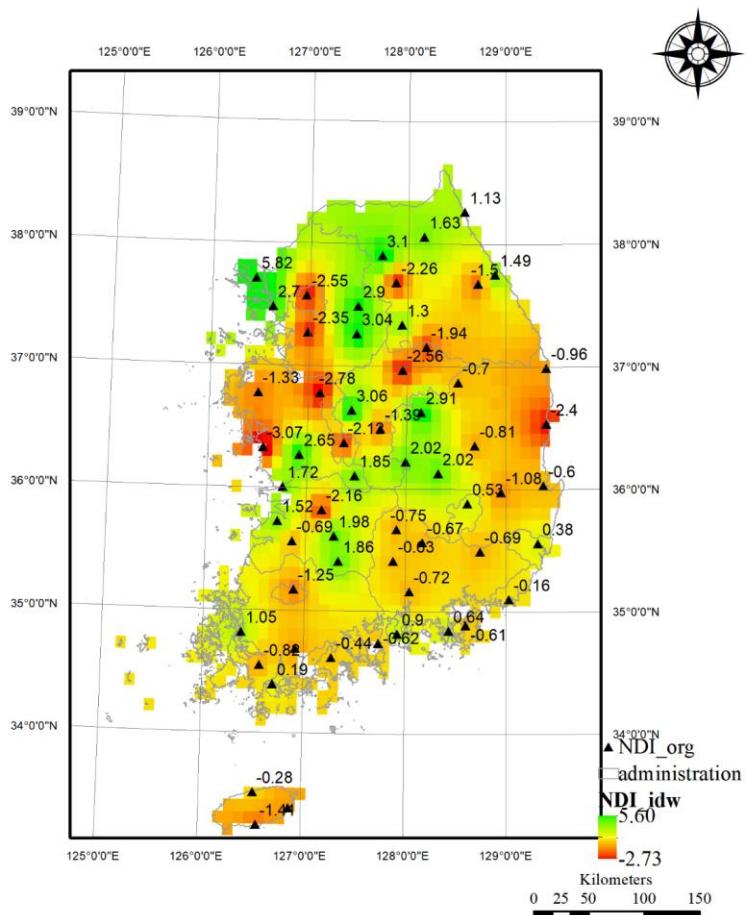


Fig 86. The interpolated drought map using IDW method.

To improve spatial drought coverage, we computed the NDI at the points do not exist in the original NDI points. These predicted points are obtained from a multiple regression ANN follows the progression.

First, we retrieve satellite-based data have related to NDI such as precipitation (p), runoff (r), soil moisture (s), evapotranspiration (e), temperature (t), wind speed (w), topographical elevation (z).

Then extracted values at original NDI to build a regression model:

$$NDI \sim p + r + s + e + t + w + x + y + z \quad (1)$$

The data set of ANN regression model is presented in the Table 1.

*Table 18. Data set for the ANN regression model.*

<b>NDI</b>	<b>p</b>	<b>r</b>	<b>s</b>	<b>e</b>	<b>t</b>	<b>w</b>	<b>x</b>	<b>y</b>	<b>z</b>
1.13	53.70	8.80E-06	0.41	2.50E-06	293.22	3.84	128.57	38.25	18.10
-1.50	52.68	3.50E-06	0.39	4.27E-05	288.44	4.21	128.72	37.68	772.40
3.10	64.02	5.50E-06	0.40	2.10E-06	292.32	3.96	127.74	37.92	75.64
1.49	53.68	1.20E-06	0.34	4.01E-05	292.77	4.24	128.89	37.75	26.04
-2.55	42.37	1.00E-06	0.40	8.00E-07	294.12	4.42	126.97	37.57	85.80
2.70	40.43	1.90E-06	0.40	4.00E-07	294.58	4.48	126.63	37.48	71.43
1.30	41.30	3.20E-06	0.40	2.00E-06	293.12	4.21	127.95	37.34	148.60
-2.35	36.50	4.00E-07	0.40	5.00E-07	294.44	4.39	126.99	37.27	34.10
-2.56	38.50	1.30E-06	0.29	3.70E-05	293.18	4.19	127.95	36.97	116.29
-1.33	32.19	1.00E-06	0.21	2.53E-05	294.91	4.40	126.49	36.78	28.91
-0.96	71.09	3.70E-06	0.35	4.09E-05	292.64	5.51	129.41	36.99	50.00
3.06	31.77	9.60E-06	0.41	2.90E-06	294.52	4.48	127.44	36.63	58.70
-2.12	35.63	9.30E-06	0.41	3.20E-06	294.42	4.43	127.37	36.37	68.90
2.02	47.15	2.10E-06	0.31	3.80E-05	292.33	4.10	127.99	36.22	243.70
-0.60	101.06	4.36E-05	0.41	4.20E-06	293.72	6.16	129.38	36.03	2.91
1.72	40.88	2.10E-06	0.21	2.61E-05	294.95	4.21	126.76	36.00	23.20
0.53	91.65	2.16E-05	0.41	3.70E-06	293.17	4.84	128.62	35.89	64.08
-2.16	49.39	9.60E-06	0.40	1.90E-06	294.55	4.21	127.12	35.84	61.40
0.38	138.69	3.02E-05	0.41	5.10E-06	293.96	6.76	129.32	35.56	33.60
-1.25	77.95	1.37E-05	0.41	2.20E-06	294.27	4.33	126.89	35.17	70.35
-0.16	139.39	1.91E-05	0.41	4.90E-06	293.72	6.64	129.03	35.10	69.60
0.64	138.85	2.39E-05	0.41	3.60E-06	294.31	6.09	128.44	34.85	32.30
1.05	84.70	4.50E-06	0.31	3.50E-05	294.53	4.53	126.38	34.82	37.99
-0.62	109.05	2.94E-05	0.41	3.60E-06	294.40	5.56	127.74	34.74	64.64
0.19	116.18	8.50E-06	0.36	4.74E-05	294.61	5.72	126.70	34.40	35.45

<b>NDI</b>	<b>p</b>	<b>r</b>	<b>s</b>	<b>e</b>	<b>t</b>	<b>w</b>	<b>x</b>	<b>y</b>	<b>z</b>
-0.28	212.70	1.87E-05	0.36	4.53E-05	293.28	6.80	126.53	33.51	20.45
-0.72	109.03	3.42E-05	0.41	3.30E-06	293.83	4.63	128.04	35.16	30.20
5.82	58.83	6.00E-07	0.22	2.64E-05	293.74	4.34	126.45	37.70	47.01
2.90	37.42	1.70E-06	0.29	3.82E-05	293.11	4.26	127.49	37.49	47.91
3.04	34.97	3.40E-06	0.23	3.24E-05	294.07	4.30	127.48	37.26	78.01
1.63	59.87	1.50E-06	0.32	3.34E-05	290.73	3.87	128.17	38.06	200.15
-2.26	48.30	5.00E-07	0.31	3.99E-05	291.93	4.12	127.88	37.70	140.91
-1.94	41.50	9.00E-07	0.31	3.99E-05	292.11	4.19	128.19	37.16	259.80
-1.39	35.93	2.20E-06	0.31	3.91E-05	293.30	4.28	127.73	36.49	175.00
-2.78	33.16	1.10E-06	0.21	2.48E-05	294.21	4.35	127.29	36.76	81.50
-3.07	33.23	6.00E-07	0.23	3.02E-05	294.92	4.27	126.56	36.33	16.90
2.65	37.49	8.00E-07	0.28	3.82E-05	294.77	4.29	126.92	36.27	11.79
1.85	42.10	2.50E-06	0.30	3.86E-05	293.31	4.48	127.48	36.10	170.34
1.52	50.49	2.40E-06	0.23	3.30E-05	294.88	4.10	126.72	35.73	11.96
1.98	57.01	2.30E-06	0.32	4.23E-05	292.98	4.14	127.29	35.61	247.11
-0.69	57.60	1.23E-05	0.40	1.40E-06	294.24	4.14	126.87	35.56	44.61
1.86	66.96	4.30E-06	0.34	4.53E-05	293.56	4.17	127.40	35.42	132.52
-0.95	91.49	6.10E-06	0.35	4.68E-05	293.86	5.23	126.92	34.69	45.02
-0.82	96.82	5.40E-06	0.29	4.08E-05	294.28	5.12	126.57	34.55	13.21
-0.44	93.56	4.30E-06	0.36	5.02E-05	294.56	5.44	127.28	34.63	53.76
-0.70	47.82	1.80E-06	0.27	3.66E-05	292.38	4.48	128.52	36.87	210.61
2.91	39.90	8.00E-07	0.30	3.63E-05	292.70	4.22	128.15	36.63	170.61
-2.40	75.28	5.60E-06	0.34	4.15E-05	292.99	5.84	129.41	36.53	40.61
-0.81	51.36	1.90E-06	0.32	4.03E-05	292.79	5.03	128.71	36.57	140.09
-0.81	57.00	1.60E-06	0.32	4.04E-05	293.00	4.81	128.69	36.36	81.80
2.02	60.22	1.45E-05	0.41	3.10E-06	293.58	4.27	128.32	36.13	48.88
-1.08	96.93	2.53E-05	0.41	3.80E-06	293.24	5.36	128.95	35.98	93.86
-0.75	76.69	4.90E-06	0.36	4.55E-05	292.16	3.91	127.91	35.68	230.20

<b>NDI</b>	<b>p</b>	<b>r</b>	<b>s</b>	<b>e</b>	<b>t</b>	<b>w</b>	<b>x</b>	<b>y</b>	<b>z</b>
-0.67	93.62	3.80E-06	0.36	4.61E-05	293.40	4.12	128.17	35.57	32.00
-0.69	123.84	2.16E-05	0.41	3.60E-06	293.31	5.48	128.74	35.49	11.20
-0.63	87.25	9.10E-06	0.36	4.78E-05	292.78	4.03	127.88	35.41	138.10
-0.61	142.88	1.08E-05	0.39	5.32E-05	293.59	6.52	128.60	34.89	45.40
0.90	114.59	9.90E-06	0.37	5.00E-05	294.22	5.34	127.93	34.82	43.70
-1.41	221.87	2.77E-05	0.41	2.60E-06	295.30	6.80	126.57	33.25	47.03

We split randomly dataset into training and testing follows the ratio 80% and 20%. The training and testing point to build the ANN model is presented in Fig 4.

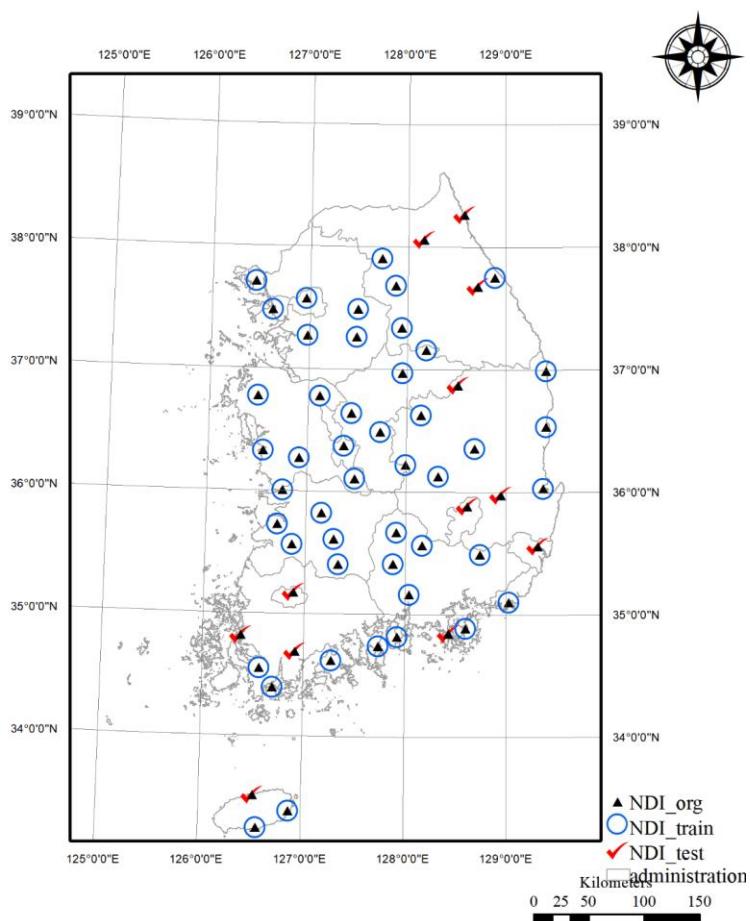


Fig 87. The location of the original NDI, random splitting training and testing data set of the ANN model.

After success builds the ANN model, we use it to predict the rest of NDI in the other locations. In this case, ANN model is applied as a regression function. Finally, using this function, we computed the rest of points do not exist in original NDI. These locations have the input values extracting from satellite-based data (The small green dots in Fig 5). The *first map* is built from ANN model.

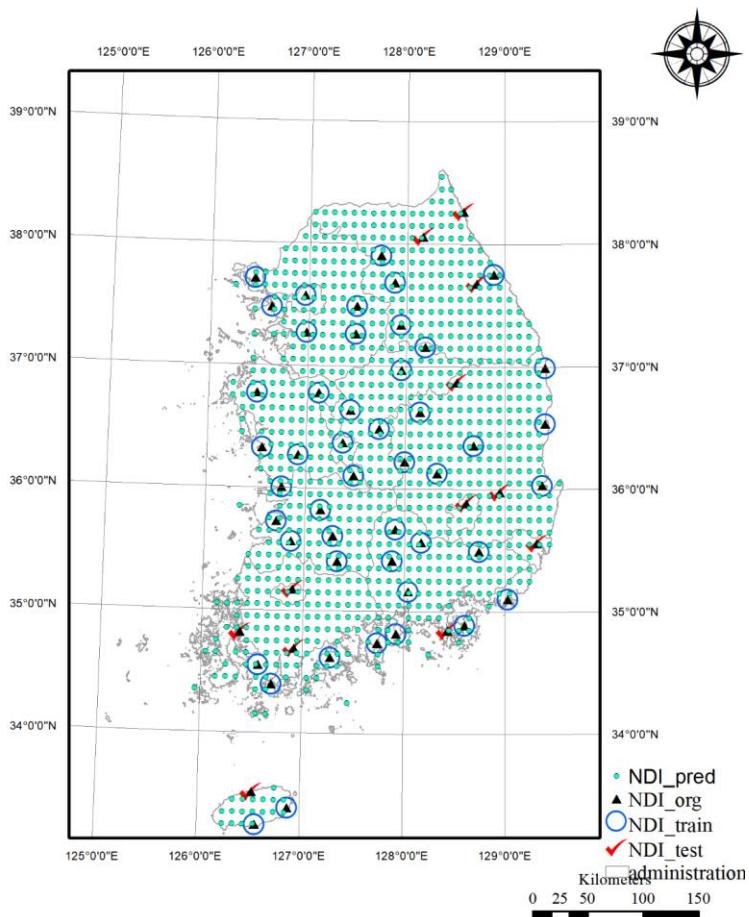


Fig 88. The location of original NDI, random splitting training, testing data set of ANN model and predicted NDI.

To estimate the accuracy of model we compare original NDI and predicted NDI from ANN model at the testing locations. To avoid overfitting, we use 10 cross validation method to evaluate the model. It means that location of training and testing are chosen randomly in 10 times. It makes sure that the training and testing cover all of the dataset. This method is used for the Kriging and cokriging model in the next parts.

Although ANN is suitable for adding the spatial characteristic, its accuracy is lower than sparse original NDI. Therefore, we propose use Ordinary Kriging to interpolate the original NDI to obtain the second map (Fig 6).

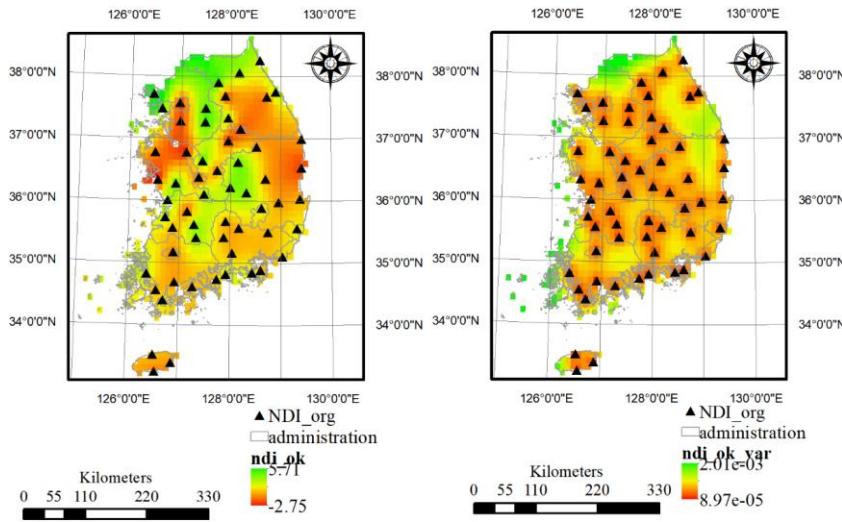


Fig 89. Predicted NDI uses OK (left) and variance of the predicted map (right).

The OK interpolated method gives a reasonable map compare to IDW. It is avoiding of “island” problem and give the variance estimation. The lower the kriging variance is closer the original NDI or the area have the highest density of original NDI points. The map is constructed from OK retraining the values at these original NDI locations. It is named as the *second map*.

To take advantage from first map and second map, we used the cokriging to merging the conditional map of NDI. Cokriging used the second map as primarily map and the first map as secondary data. Try to keep both mean and minimise the variance is main achievement of cokriging. Fig 7 present predicted NDI map from ANN, OK, and cokriging model.

Again, cross validation is used to estimate the accuracy of model at original NDI locations. The errors at 59 original NDI location is used as the validation data for all models.

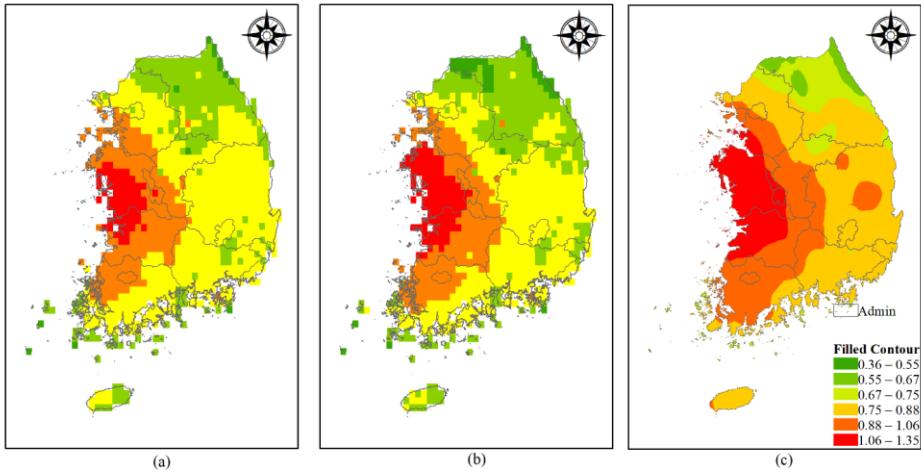


Fig 90. Compares the predicted NDI map: (a) from ANN, (b) from OK, and (c) from cokriging.

## 5. Discussion

(6) In this study, several spatial interpolation methods were analyzed. The deterministic method such as Inverse Distance Weight (IDW) is simple, easy to apply for interpolation. But it does not include the spatial autocorrelation of whole data. The geostatistical method such as Kriging considered the spatial autocorrelation, the variance for predicted map. However, it is complicated. It requires optimized parameters of the model, chose the best fit variogram that is difficult. The machine learning such as ANN is also the promising approach for spatial analysis. The drawback of ANN is difficult to explain the correlation of predictors and predicted values. Therefore, one method cannot fit all. The combination of geostatistic and machine learning could improve the accuracy of spatial interpolation, or other spatial downscaling.

(7) One of the most challenges of spatial study is limited observational data for model evaluation. To overcome this issue, we proposed to use Leave-one-out cross-validation (LOOC). LOOC is a

special case of cross-validation. The number of folds equals the number of instances in the data set. Thus, the learning algorithm is applied once for each instance, using all other instances as a training set and using the selected instance as a single-item test set. This process is closely related to the statistical method of jack-knife estimation (Efron, 1982). In our study, original NDI at 59 ASOS is considered as a sample. We random keep 1 station out of sample. The rest of 58/59 original NDI are used as the input for setup ANN model, and KR model. In the corkring model, we used K-Fold cross validation because the final map is merging both ANN and KR model. The input data are diversified.

## References

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*: SIAM.

## WEEK 31

**A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network**

### 6. Clarify these unclear parts of the methodology

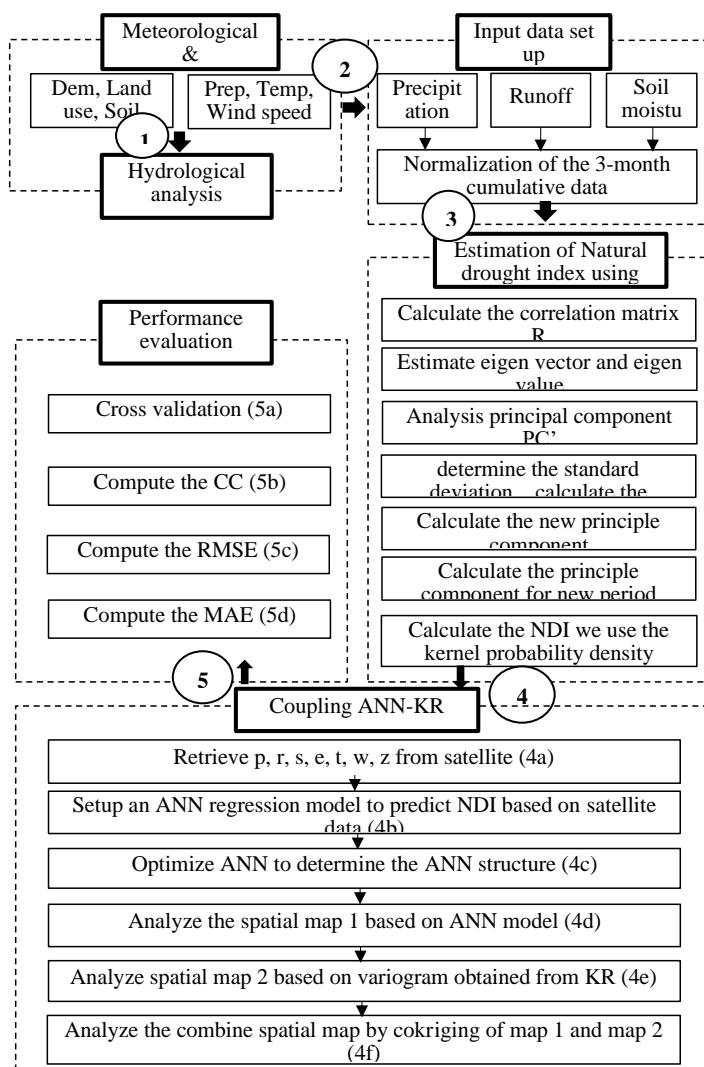
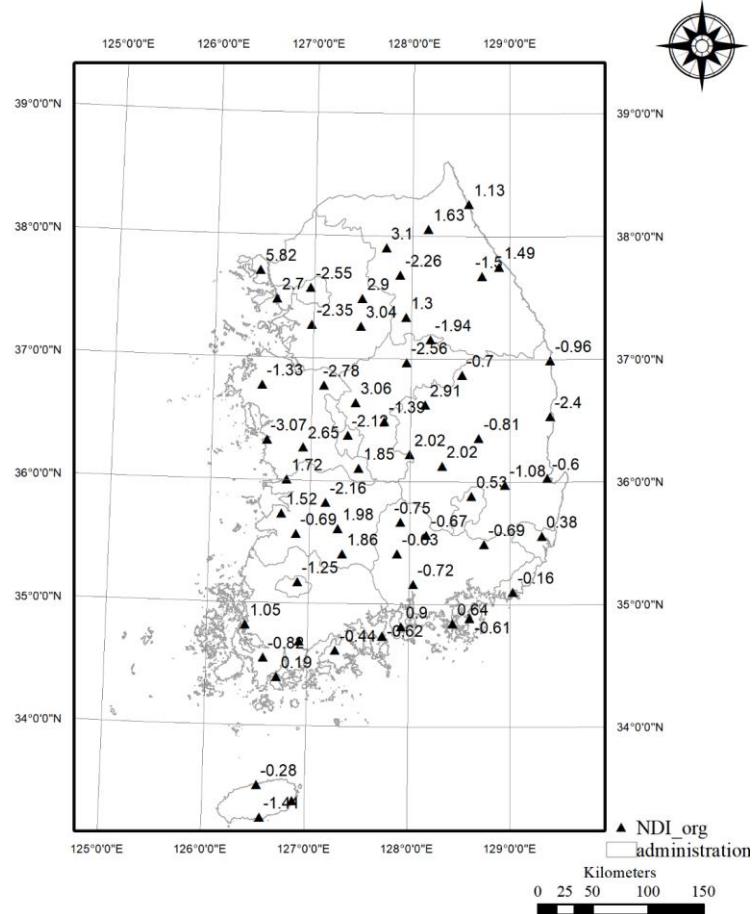


Fig 91.

Complete step 3, we got the extreme drought at 59 ASOS stations in August 2015 (Fig 2).



*Fig 92. The original NDI computed from precipitation, runoff, soil moisture.*

In the previous study, we used the invert distance weight (IDW) to interpolate the spatial drought map. Results from IDW depend on the distance of original NDI points and predicted NDI locations. The predicted points are closer the original points are better estimation. While the other predicted points that are far from original points are difficult to predict accurately. As the name, map from IDW has a trend to create “islands” with centroid are the original points (Fig 3). Furthermore, IDW does not consider the spatial autocorrelation of predicted map.

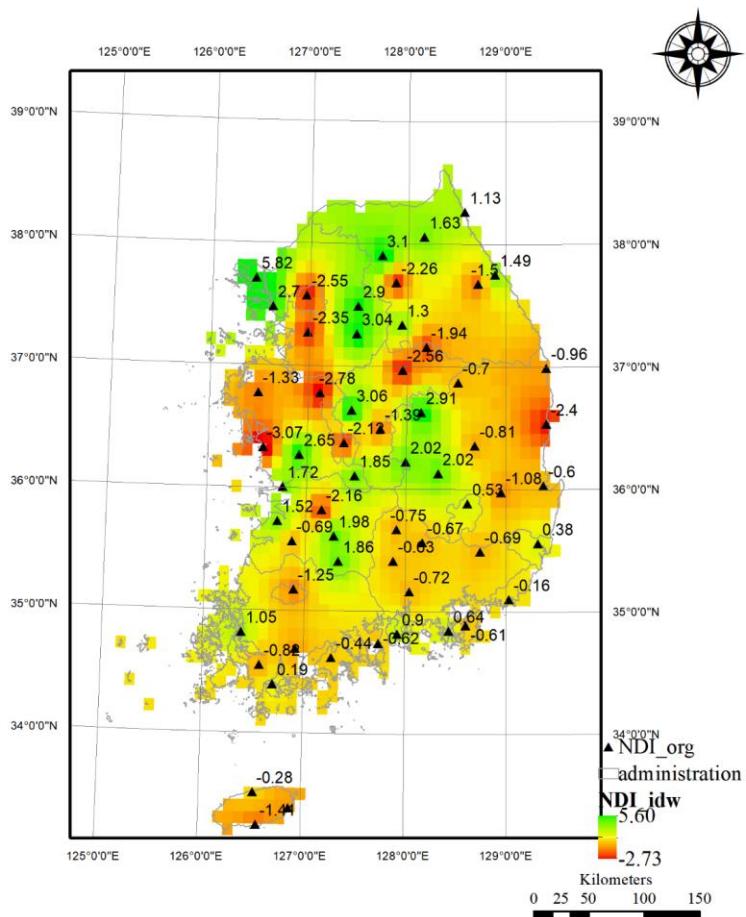


Fig 93. The interpolated drought map using IDW method.

To improve spatial drought coverage, we computed the NDI at the points do not exist in the original NDI points. These predicted points are obtained from a multiple regression ANN follows the progression.

First, we retrieve satellite-based data have related to NDI such as precipitation (p), runoff (r), soil moisture (s), evapotranspiration (e), temperature (t), wind speed (w), topographical elevation (z).

Then extracted values at original NDI to build a regression model:

$$NDI \sim p + r + s + e + t + w + x + y + z \quad (1)$$

The data set of ANN regression model is presented in the Table 1.

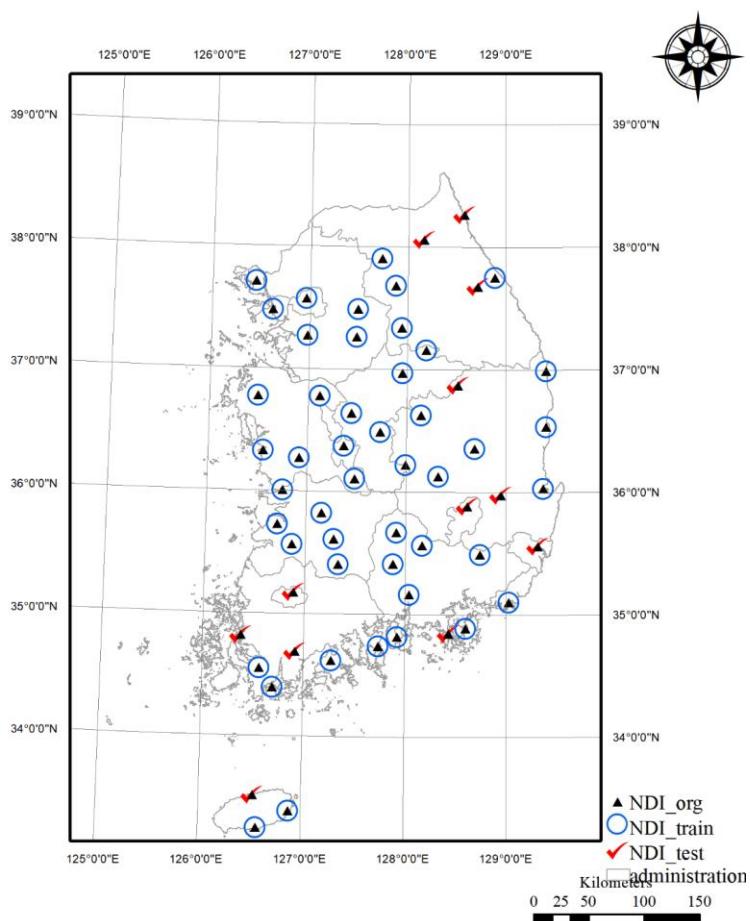
*Table 19. Data set for the ANN regression model.*

<b>NDI</b>	<b>p</b>	<b>r</b>	<b>s</b>	<b>e</b>	<b>t</b>	<b>w</b>	<b>x</b>	<b>y</b>	<b>z</b>
1.13	53.70	8.80E-06	0.41	2.50E-06	293.22	3.84	128.57	38.25	18.10
-1.50	52.68	3.50E-06	0.39	4.27E-05	288.44	4.21	128.72	37.68	772.40
3.10	64.02	5.50E-06	0.40	2.10E-06	292.32	3.96	127.74	37.92	75.64
1.49	53.68	1.20E-06	0.34	4.01E-05	292.77	4.24	128.89	37.75	26.04
-2.55	42.37	1.00E-06	0.40	8.00E-07	294.12	4.42	126.97	37.57	85.80
2.70	40.43	1.90E-06	0.40	4.00E-07	294.58	4.48	126.63	37.48	71.43
1.30	41.30	3.20E-06	0.40	2.00E-06	293.12	4.21	127.95	37.34	148.60
-2.35	36.50	4.00E-07	0.40	5.00E-07	294.44	4.39	126.99	37.27	34.10
-2.56	38.50	1.30E-06	0.29	3.70E-05	293.18	4.19	127.95	36.97	116.29
-1.33	32.19	1.00E-06	0.21	2.53E-05	294.91	4.40	126.49	36.78	28.91
-0.96	71.09	3.70E-06	0.35	4.09E-05	292.64	5.51	129.41	36.99	50.00
3.06	31.77	9.60E-06	0.41	2.90E-06	294.52	4.48	127.44	36.63	58.70
-2.12	35.63	9.30E-06	0.41	3.20E-06	294.42	4.43	127.37	36.37	68.90
2.02	47.15	2.10E-06	0.31	3.80E-05	292.33	4.10	127.99	36.22	243.70
-0.60	101.06	4.36E-05	0.41	4.20E-06	293.72	6.16	129.38	36.03	2.91
1.72	40.88	2.10E-06	0.21	2.61E-05	294.95	4.21	126.76	36.00	23.20
0.53	91.65	2.16E-05	0.41	3.70E-06	293.17	4.84	128.62	35.89	64.08
-2.16	49.39	9.60E-06	0.40	1.90E-06	294.55	4.21	127.12	35.84	61.40
0.38	138.69	3.02E-05	0.41	5.10E-06	293.96	6.76	129.32	35.56	33.60
-1.25	77.95	1.37E-05	0.41	2.20E-06	294.27	4.33	126.89	35.17	70.35
-0.16	139.39	1.91E-05	0.41	4.90E-06	293.72	6.64	129.03	35.10	69.60
0.64	138.85	2.39E-05	0.41	3.60E-06	294.31	6.09	128.44	34.85	32.30
1.05	84.70	4.50E-06	0.31	3.50E-05	294.53	4.53	126.38	34.82	37.99
-0.62	109.05	2.94E-05	0.41	3.60E-06	294.40	5.56	127.74	34.74	64.64
0.19	116.18	8.50E-06	0.36	4.74E-05	294.61	5.72	126.70	34.40	35.45

<b>NDI</b>	<b>p</b>	<b>r</b>	<b>s</b>	<b>e</b>	<b>t</b>	<b>w</b>	<b>x</b>	<b>y</b>	<b>z</b>
-0.28	212.70	1.87E-05	0.36	4.53E-05	293.28	6.80	126.53	33.51	20.45
-0.72	109.03	3.42E-05	0.41	3.30E-06	293.83	4.63	128.04	35.16	30.20
5.82	58.83	6.00E-07	0.22	2.64E-05	293.74	4.34	126.45	37.70	47.01
2.90	37.42	1.70E-06	0.29	3.82E-05	293.11	4.26	127.49	37.49	47.91
3.04	34.97	3.40E-06	0.23	3.24E-05	294.07	4.30	127.48	37.26	78.01
1.63	59.87	1.50E-06	0.32	3.34E-05	290.73	3.87	128.17	38.06	200.15
-2.26	48.30	5.00E-07	0.31	3.99E-05	291.93	4.12	127.88	37.70	140.91
-1.94	41.50	9.00E-07	0.31	3.99E-05	292.11	4.19	128.19	37.16	259.80
-1.39	35.93	2.20E-06	0.31	3.91E-05	293.30	4.28	127.73	36.49	175.00
-2.78	33.16	1.10E-06	0.21	2.48E-05	294.21	4.35	127.29	36.76	81.50
-3.07	33.23	6.00E-07	0.23	3.02E-05	294.92	4.27	126.56	36.33	16.90
2.65	37.49	8.00E-07	0.28	3.82E-05	294.77	4.29	126.92	36.27	11.79
1.85	42.10	2.50E-06	0.30	3.86E-05	293.31	4.48	127.48	36.10	170.34
1.52	50.49	2.40E-06	0.23	3.30E-05	294.88	4.10	126.72	35.73	11.96
1.98	57.01	2.30E-06	0.32	4.23E-05	292.98	4.14	127.29	35.61	247.11
-0.69	57.60	1.23E-05	0.40	1.40E-06	294.24	4.14	126.87	35.56	44.61
1.86	66.96	4.30E-06	0.34	4.53E-05	293.56	4.17	127.40	35.42	132.52
-0.95	91.49	6.10E-06	0.35	4.68E-05	293.86	5.23	126.92	34.69	45.02
-0.82	96.82	5.40E-06	0.29	4.08E-05	294.28	5.12	126.57	34.55	13.21
-0.44	93.56	4.30E-06	0.36	5.02E-05	294.56	5.44	127.28	34.63	53.76
-0.70	47.82	1.80E-06	0.27	3.66E-05	292.38	4.48	128.52	36.87	210.61
2.91	39.90	8.00E-07	0.30	3.63E-05	292.70	4.22	128.15	36.63	170.61
-2.40	75.28	5.60E-06	0.34	4.15E-05	292.99	5.84	129.41	36.53	40.61
-0.81	51.36	1.90E-06	0.32	4.03E-05	292.79	5.03	128.71	36.57	140.09
-0.81	57.00	1.60E-06	0.32	4.04E-05	293.00	4.81	128.69	36.36	81.80
2.02	60.22	1.45E-05	0.41	3.10E-06	293.58	4.27	128.32	36.13	48.88
-1.08	96.93	2.53E-05	0.41	3.80E-06	293.24	5.36	128.95	35.98	93.86
-0.75	76.69	4.90E-06	0.36	4.55E-05	292.16	3.91	127.91	35.68	230.20

<b>NDI</b>	<b>p</b>	<b>r</b>	<b>s</b>	<b>e</b>	<b>t</b>	<b>w</b>	<b>x</b>	<b>y</b>	<b>z</b>
-0.67	93.62	3.80E-06	0.36	4.61E-05	293.40	4.12	128.17	35.57	32.00
-0.69	123.84	2.16E-05	0.41	3.60E-06	293.31	5.48	128.74	35.49	11.20
-0.63	87.25	9.10E-06	0.36	4.78E-05	292.78	4.03	127.88	35.41	138.10
-0.61	142.88	1.08E-05	0.39	5.32E-05	293.59	6.52	128.60	34.89	45.40
0.90	114.59	9.90E-06	0.37	5.00E-05	294.22	5.34	127.93	34.82	43.70
-1.41	221.87	2.77E-05	0.41	2.60E-06	295.30	6.80	126.57	33.25	47.03

We split randomly dataset into training and testing follows the ratio 80% and 20%. The training and testing point to build the ANN model is presented in Fig 4.



*Fig 94. The location of the original NDI, random splitting training and testing data set of the ANN model.*

After success builds the ANN model, we use it to predict the rest of NDI in the other locations. In this case, ANN model is applied as a regression function. Finally, using this function, we computed the rest of points do not exist in original NDI. These locations have the input values extracting from satellite-based data (The small green dots in Fig 5). The **first map** is built from ANN model.

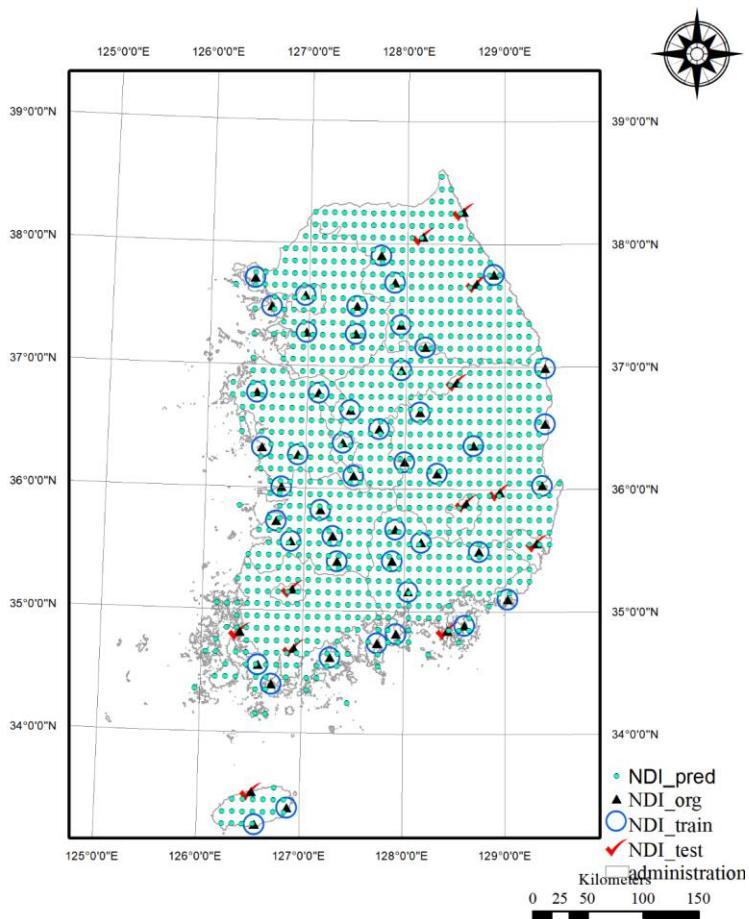


Fig 95. The location of original NDI, random splitting training, testing data set of ANN model and predicted NDI.

To estimate the accuracy of model we compare original NDI and predicted NDI from ANN model at the testing locations. To avoid overfitting, we use 10 cross validation method to evaluate the model. It means that location of training and testing are chosen randomly in 10 times. It makes sure that the training and testing cover all of the dataset. This method is used for the Kriging and cokriging model in the next parts.

Although ANN is suitable for adding the spatial characteristic, its accuracy is lower than sparse original NDI. Therefore, we propose use Ordinary Kriging to interpolate the original NDI to obtain the second map (Fig 6).

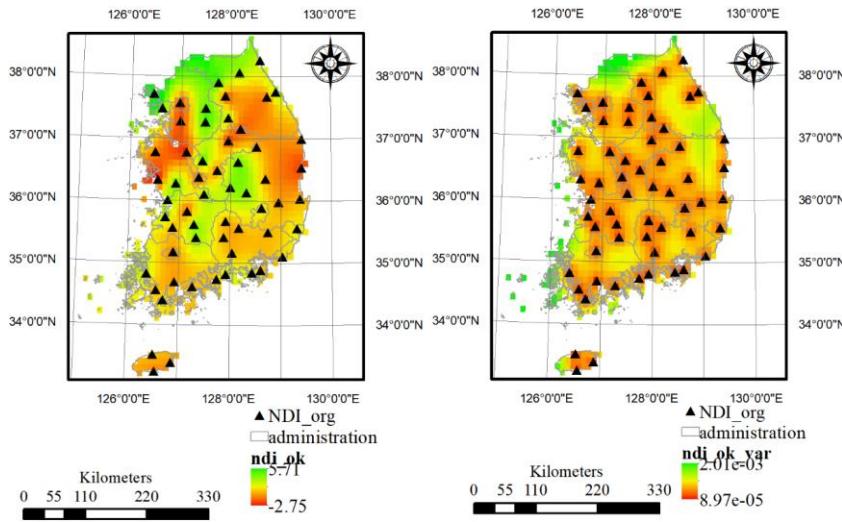


Fig 96. Predicted NDI uses OK (left) and variance of the predicted map (right).

The OK interpolated method gives a reasonable map compare to IDW. It is avoiding of “island” problem and give the variance estimation. The lower the kriging variance is closer the original NDI or the area have the highest density of original NDI points. The map is constructed from OK retraining the values at these original NDI locations. It is named as the *second map*.

To take advantage from first map and second map, we used the cokriging to merging the conditional map of NDI. Cokriging used the second map as primarily map and the first map as secondary data. Try to keep both mean and minimise the variance is main achievement of cokriging. Fig 7 present predicted NDI map from ANN, OK, and cokriging model.

Again, cross validation is used to estimate the accuracy of model at original NDI locations. The errors at 59 original NDI location is used as the validation data for all models.

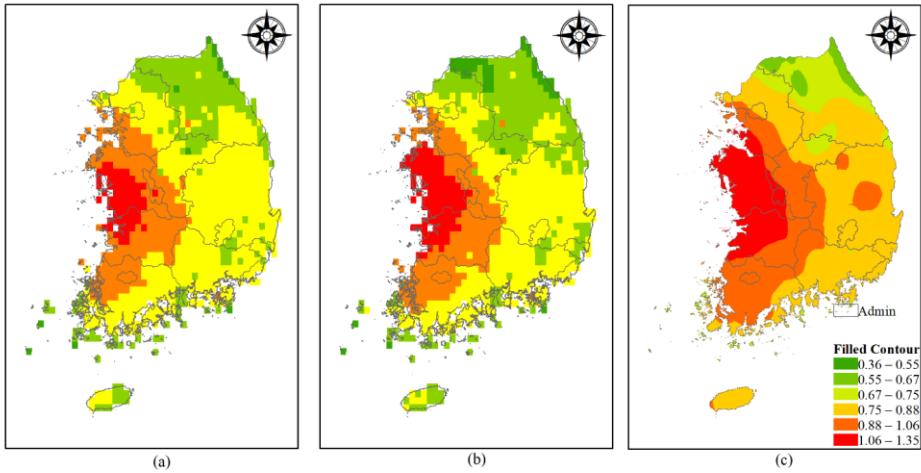


Fig 97. Compares the predicted NDI map: (a) from ANN, (b) from OK, and (c) from cokriging.

## 7. Discussion

(8) In this study, several spatial interpolation methods were analyzed. The deterministic method such as Inverse Distance Weight (IDW) is simple, easy to apply for interpolation. But it does not include the spatial autocorrelation of whole data. The geostatistical method such as Kriging considered the spatial autocorrelation, the variance for predicted map. However, it is complicated. It requires optimized parameters of the model, chose the best fit variogram that is difficult. The machine learning such as ANN is also the promising approach for spatial analysis. The drawback of ANN is difficult to explain the correlation of predictors and predicted values. Therefore, one method cannot fit all. The combination of geostatistic and machine learning could improve the accuracy of spatial interpolation, or other spatial downscaling.

(9) One of the most challenges of spatial study is limited observational data for model evaluation. To overcome this issue, we proposed to use Leave-one-out cross-validation (LOOC). LOOC is a

special case of cross-validation. The number of folds equals the number of instances in the data set. Thus, the learning algorithm is applied once for each instance, using all other instances as a training set and using the selected instance as a single-item test set. This process is closely related to the statistical method of jack-knife estimation (Efron, 1982). In our study, original NDI at 59 ASOS is considered as a sample. We random keep 1 station out of sample. The rest of 58/59 original NDI are used as the input for setup ANN model, and KR model. In the corkring model, we used K-Fold cross validation because the final map is merging both ANN and KR model. The input data are diversified.

## References

Efron, B. (1982). *The jackknife, the bootstrap and other resampling plans*: SIAM.

## WEEK 32

**A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network**

### 8. Checking analyzed results

#### *|Checking data input*

In this study, original NDI was computed by using precipitation, runoff and soil moisture is the first type of input data. It is named as GBD (Gauge-Based Data). Like the most popular drought index-SPI, the NDI has different temporal scale: 1 month, 3 months, 6 months, 12 months, 24 months. Timescale 1, 3, 6 months is used for short term estimation (Bae et al., 2018; Kim et al., 2011; Son et al., 2011). Timescale such as 12 months is used for long term evaluation (Lee et al., 2019) . Timescale should be determined. We determine to use 3 months temporal scale for our study as short-term estimation. Because the main of our study is estimated spatial extreme drought. The extreme drought event occurs in the short term. Therefore, 3 months-time scale is suitable for our study. The addition of NDI should be considered that is the difference between projected coordinate system of NDI and Global projected coordinate system of SBD. We should check the consistency of these coordinate systems.

The second type data is retrieved from satellite data. It consists of precipitation (p), runoff (r), soil moisture (s), evapotranspiration (e), temperature (t), average wind speed (w). It is named as SBD (Satellite-Based-Data). The units of SBD can lead to confusion. For instance, units of evapotranspiration from GLDAS is  $kg \cdot m^{-2} \cdot s^{-1}$  . We need to convert to monthly by a multiple of 3600 (number of second for 1 hour) (24 (number of hours for 1 day) (number of the days correspond to that month. The unit of runoff from GLDAS in  $kg \cdot m^{-2}$  Accumulated over 3-hour interval. To convert to monthly, it needs to multiple of 8 (3 hours/ day) (number of days in that month. Table 1 presents some important parameters should be checked.

Table 20. List of checked parameters of input data.

No.	Data	Parameters	Notes
1	NDI	-Timescale: 3 months - Temporal resolution: monthly -Spatial resolution: ~190 km -Coordinate systems	Corrected
2	Precipitation	-Sources: IMERG -Spatial resolution: 0.1 degrees -Temporal resolution: monthly -Unit: mm/month	Corrected
3	Runoff	-Sources: FDAS NOAH01_C_GL_Mav001 Spatial resolution: 0.1 degrees -Temporal resolution: monthly -Unit: $kg \cdot m^{-2} \cdot s^{-1}$	Uncorrected
4	Soil moisture	-Sources: FDAS NOAH01_C_GL_Mav001 -Spatial resolution: 0.1 degrees -Temporal resolution: monthly -Unit: $m^3 \cdot m^{-3}$	Corrected
5	Evapotranspiration	-Sources: FDAS NOAH01_C_GL_Mav001 -Spatial resolution: 0.1 degrees -Temporal resolution: monthly -Unit: $kg \cdot m^{-2} \cdot s^{-1}$	Uncorrected
6	Temperature	-Sources: FDAS NOAH01_C_GL_Mav001  -Spatial resolution: 0.1 degrees -Temporal resolution: monthly -Unit: Kelvin	Corrected
7	Average wind speed	-Sources: FDAS NOAH01_C_GL_Mav001 -Spatial resolution: 0.1 degrees -Temporal resolution: monthly -Unit: m/s	Corrected

Fig 1 presents the results before correction data units. Before building an ANN model, all data were standardized in the range from 0 to 1. They also removed dimension. Therefore, the unit does

not impact the training and prediction results. But please note that results from ANN must be rescaled to original dimension follow equation:

$$NDI_{org} = NDI_{model} \times (NDImax - NDImin) + NDImin \quad (1)$$

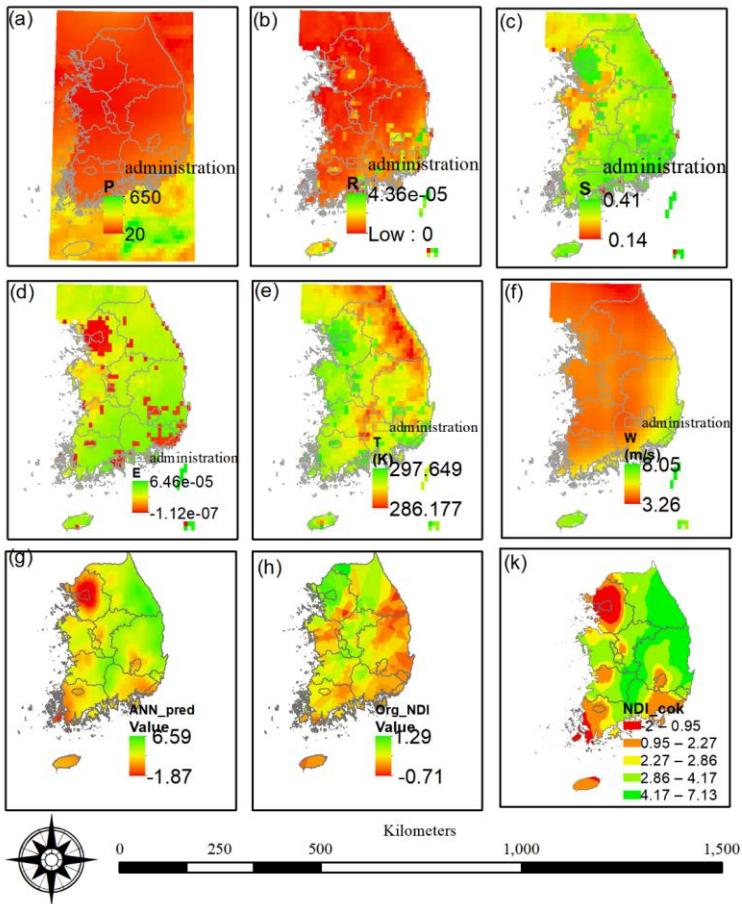


Fig 98.The SDB input and results: (a) precipitation, (b) runoff, (c) soil moisture, (d) evapotranspiration, (e) temperature, (f) average wind speed, (g) NDI prediction from ANN, (h) NDI prediction from OK, (k) NDI prediction from COK.

We corrected units for runoff, evapotranspiration follow equation:

$$X = X_{satellite} \times 3600 \times 24 \times 30 \quad (2)$$

The results predicted from ANN and Cokring are different to expectation (Fig 2).

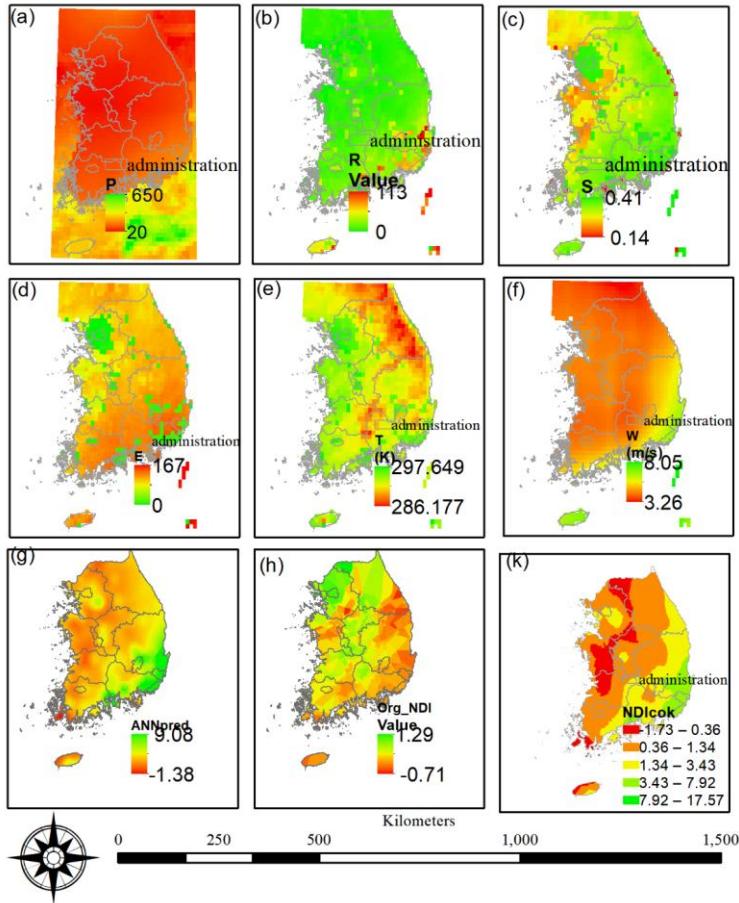
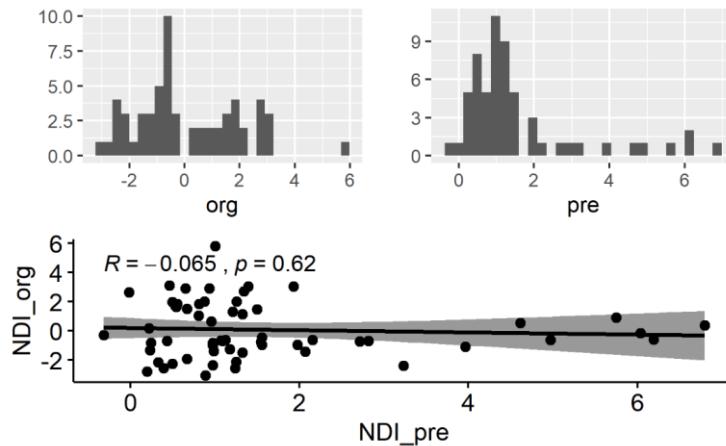


Fig 99. The SDB input and results after corrected units: (a) precipitation, (b) runoff, (c) soil moisture, (d) evapotranspiration, (e) temperature, (f) average wind speed, (g) NDI prediction from ANN, (h) NDI prediction from OK, (k) NDI prediction from COK.

The extreme drought area in the north of South Korea has disappeared in both ANN prediction and COK model. The Easter of study becomes wet compare to before corrected units.

It is difficult to compare prediction for the whole region. We extracted NDI values at 59 ASOS from model prediction. Then is used to compare to original NDI.

In Fig 3 present the comparison of original NDI (org), predicted NDI from ANN (pre), and correlation of them.



*Fig 100. Histogram of original NDI, predicted NDI, and correlation.*

The histogram from the org and pre is various, and the Pearson correlation is low ( $0.065$ ), and it is not significant ( $p = 0.62 \geq 5\%$ ). Therefore, results of ANN do not fit to original NDI.

|*Checking optimal model*

|*Checking evaluation*

## 9. Discussion

(10) After checking and correction units of satellite-based data, model gives the diverse results. It can be a cause of filling up in the input data. We filled the missing data by using mean value. It leads to being inaccuracy for input model. In addition, checking optimal model and evaluation were not examined. Therefore, it is hard to make a conclusion of our model.

(11) Checking analyzed results is very important to confirm the reasonable output. We suggest spending more time to criticize the computation and set up the model. First, we should figure out what are the key characters in our model. The key characters that impacted to result seriously. Then the analysis sensitive model should be considered as needed.

## References

- Bae, S., Lee, S.-H., Yoo, S.-H., & Kim, T. (2018). Analysis of Drought Intensity and Trends Using the Modified SPEI in South Korea from 1981 to 2010. *Water*, 10(3), 327.
- Kim, D.-W., Byun, H.-R., Choi, K.-S., & Oh, S.-B. (2011). A Spatiotemporal Analysis of Historical Droughts in Korea. *Journal of Applied Meteorology and Climatology*, 50(9), 1895-1912. doi:10.1175/2011jamc2664.1
- Lee, M.-H., Im, E.-S., & Bae, D.-H. (2019). A comparative assessment of climate change impacts on drought over Korea based on multiple climate projections and multiple drought indices. *Climate Dynamics*, 1-16.
- Son, K.-H., Bae, D.-H., & Chung, J.-S. (2011). Drought Analysis and Assessment by Using Land Surface Model on South Korea. *Journal of Korea Water Resources Association*, 44(8), 667-681. doi:10.3741/JKWRA.2011.44.8.667

Theo ý thầy:

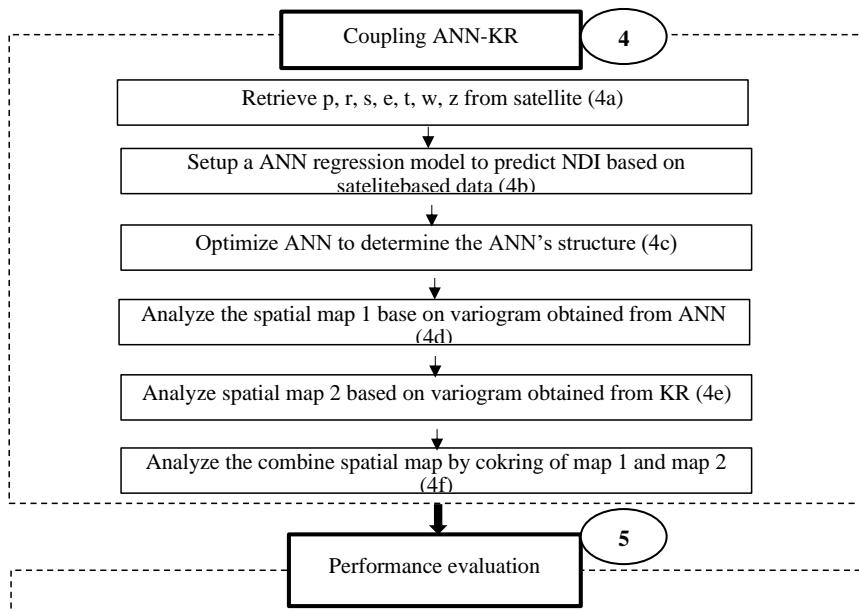
- Dùng máy học để tìm ra quy luật hỗ trợ cho việc nội suy
- Đánh giá kết quả nó tốt hơn phép nội suy IDW là được
- Đọc thêm bài của Dr. Yoong

## WEEK 33

A framework evaluates extreme drought coverage using natural drought index, Satellite based data

and artificial neural network

### 10. Understand problems in our methodology



The step 4 is not clear. It does not make sense because we can compute NDI by precipitation, runoff and soil moisture. It is better than using ANN model. The other things, we are training model for small dataset at 59 stations, but we use it for prediction of 1010 points. Therefore, the bias is very high. This leads to the un-correlation between the predicted model and the original Fig 101. Step 4 and step 5 of the spatial extreme drought coverages assessment.

## 1. Understand conditional merging method

Three key papers were reviewed to get the overview about conditional merging technique using to combine radar (satellite) and gauge data.

### |Cokriging

Firstly, we review about the basic idea using radar-rainfall and rain gage data (Frajewski, 1987). An ordinary cokriging procedure was used to merge rainfall data from radars and standard rain gages. The covariance matrices required to perform cokriging are computed from single realization data. The assume that the ground is truth and error structure of radar data are unknown. It needs to determine the covariance between radar data and the true rainfall. The advantage of the procedure to remove the bias in radar is tested.

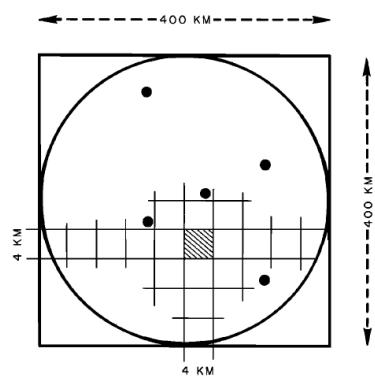


Fig 102. Schematic presentation of the radar (squares) and rain gage (dots) data in the domain  $\Omega$ .

It assumes that the precipitation measurement network covering space  $\Omega$  included a weather radar, and a set of  $N$  rain gages. The radar produces accumulated rainfall on a rectangular grid. The rain gage data present point measurement. Both data sets are depicted in Fig 2. Rain gauge data is considered to provide good point accuracy, but it offers little information on the spatial distribution. On the other hand, Radar is capable of accurately delineating rainfall, but less accuracy cause of various meteorology, equipment, methodological factor. Therefore, combine the two data sets to get high spatial and point accuracy. Mathematical equations by that method follow by:

$$R_{ij} = \frac{1}{|A_R|} \int_{A_R} Z(u_{ij}) du + \varepsilon_{R_{ij}} \quad (1)$$

$$G_k = Z(u_k) + \varepsilon_{G_k} \quad (2)$$

Where  $A_R$  is the integration area of a single radar measurement, i, j are coordinates of corresponding location,  $\varepsilon_{R_{ij}}$  is the error associated with ijth radar observation, and  $\varepsilon_{G_k}$  is the error associated with the kth gage observation.

$$V(u_0) = \frac{1}{|A|} \int_A Z(u_0) du + \varepsilon_{R_{ij}} \quad (3)$$

$$V^*(u_0) = \sum_i^{N_G} \lambda_{G_i} G_i(u_i) + \sum_i^{N_R} \lambda_{R_i} R_i(u_i) \quad (4)$$

$V^*(u_0)$ ,  $V(u_0)$  is an estimate of the mean areal precipitation on the ground level over the same area by radar.

$$\begin{aligned} \sigma_v^2 &= E[(V - V^*)^2] \\ &= \frac{1}{|A|^2} \iint_A Cov(u - v) dudv \\ &\quad - 2 \sum_{i=1}^{N_G} \lambda_{G_i} \frac{1}{|A|} \int_A cov_{GV}(u - u_i) du - 2 \sum_{j=1}^{N_G} \lambda_{R_j} \frac{1}{|A|} \int_A cov_{RV}(u - u_j) du \\ &\quad + \sum_{i=1}^{N_G} \sum_{j=1}^{N_R} \lambda_{G_i} \lambda_{R_j} Cov_{GR}(u_i - u_j) + \sum_{i=1}^{N_R} \sum_{j=1}^{N_G} \lambda_{R_i} \lambda_{G_j} Cov_{GR}(u_i - u_j) \\ &\quad + 2 \sum_{i=1}^{N_R} \sum_{j=1}^{N_R} \lambda_{R_i} \lambda_{R_j} Cov_{RR}(u_i - u_j) \end{aligned} \quad (5)$$

$$E\{V^*\} = E\{V\} \quad (6)$$

$$\sum_i^{N_G} \lambda_{G_i} = 1 \quad (7)$$

$$\sum_{j=1}^{N_R} \lambda_{R_j} = 0 \quad (8)$$

$$\sum_i^{N_G} \lambda_{G_i} + \sum_{j=1}^{N_R} \lambda_{R_j} = 1 \quad (9)$$

$$\begin{vmatrix} Cov_{RR} & Cov_{RG} & 1 & 0 \\ Cov_{GR} & Cov_{GG} & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{vmatrix} \times \begin{vmatrix} \lambda_R \\ \lambda_G \\ \mu_R \\ \mu_G \end{vmatrix} = \begin{vmatrix} Cov_{VG} \\ Cov_{VR} \\ 0 \\ 1 \end{vmatrix} \quad (10)$$

The computational algorithm can be summarized as follows:

1. Estimate the rain gage data covariance function, using the exponential isotropic model.  
This done by least square fit.
2. Block Krige the rain gage data to estimate
3. Estimate  $Cov_{RR}$ ,  $Cov_{GG}$ ,  $Cov_{RG}$  from the radar and the new kriged rain gage filed. The element of  $Cov_{RG}$  represent estimates of the cross covariance between two areal average observation fields.
4. Model  $Cov_{RR}$ ,  $Cov_{GG}$ ,  $Cov_{RG}$ , using the exponential isotropic model. This is done by least squares fit.
5. Compute  $\lambda_R$ , and  $\lambda_G$  are corresponding coefficients weight.
6. Cokrige the two fields.

To solve the equation (10), we need to estimate the vectors  $Cov_{VG}$ ,  $Cov_{VR}$ . They are covariances between radar and rain gage data, and true precipitation V. Since V is unknown,  $Cov_{VG}$  ,and  $Cov_{VR}$  is approximated:

$$Cov_{VR} = \beta_R Cov_{RR} \quad \beta_R \in (0,1) \quad (11)$$

$$Cov_{VG} = \beta_G Cov_{GG} \quad \beta_G \in (0,1) \quad (12)$$

One,  $\lambda_R$ , and  $\lambda_G$  are determined, we can calculate the variances as:

$$\hat{\sigma}_v^2 = Cov_{GG}(u_0, u_0) - \mu_G - \sum_{i=1}^{N_R} \lambda_{R_i} Cov_{VG}(u_0, u_i) \quad (13)$$

The conclusion was extracted from this paper as follow. The configuration of a precipitation-processing system is using data from radar and a network of rain gages, includes a separate bias removal procedure that is unbiased radar-rainfall filed enters the merging steps.

#### |Conditional merging

The secondly, we investigated the method combine radar and rain gauge rainfall estimation using conditional merging (Sinclair & Pegram, 2005). This method uses the Kriging to extract the optimal information content of the observed data. A mean field based on the kriged rain gauge data is adopted, while the spatial detail from radar is retained, reducing bias, but keeping the spatial variability observed by the radar. The variance of the estimation is reduced in the vicinity of gauges where they can provide good information on the true rainfall field. The conditional merging technique is suggested to use of Ordinary Kriging (Cressie, 1993) to extract the information content of the observed data. Fig 3 gives an overview of the technique.

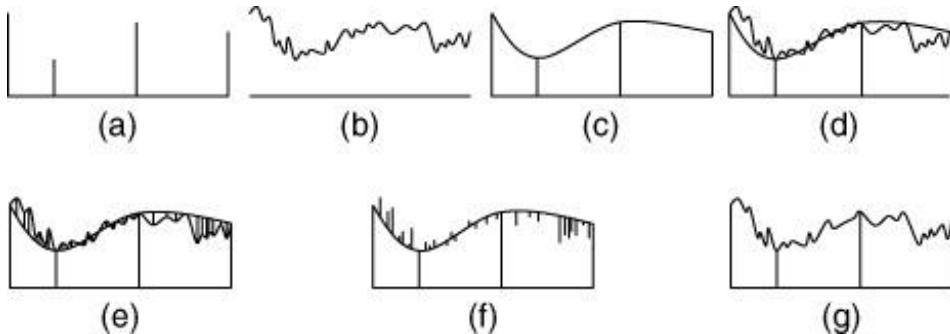


Fig 103. The conditional merging process: (a) The rainfall field is observed at discrete points by gauges, (b) the rainfall field is also observed by radar on a regular, volume-integrated grid, (c) Kriging of the rain gauge observations is used to obtain the best linear unbiased estimate of rainfall on the radar grid, (d) The radar pixel values at the rain gauge locations are interpolated onto the radar grid using Kriging, (e) At each grid point, the deviation  $C$  between the observed and interpolated radar value is computed, (f) The field of deviations obtained from (e) is applied

to the interpolated rainfall field obtained from Kriging the gauge observations, (g) A rainfall filed that follows the mean filed of rain gauge interpolation, while preserving the mean field deviations and the spatial structure of the radar field is obtained.

The error structure of the merged estimation can be examined by equations follow:

$$Z(s) = G_K(s) + \varepsilon_G(s) \quad (14)$$

$$R(s) = R_K(s) + \varepsilon_R(s) \quad (15)$$

$$M(s) = G_K(s) + \varepsilon_R(s) \quad (16)$$

$$E[Z(s) - M(s)] = E[\varepsilon_G(s) - \varepsilon_R(s)] \quad (17)$$

$$\begin{aligned} \text{var}[Z(s) - M(s)] &= \text{var}[\varepsilon_G(s) - \varepsilon_R(s)] = \sigma_{\varepsilon_G(s)}^2 + \sigma_{\varepsilon_R(s)}^2 - 2 \text{cov}[\varepsilon_G(s), \varepsilon_R(s)] \\ &= \beta - 2\sigma_{\varepsilon_G(s)}\sigma_{\varepsilon_R(s)} \end{aligned} \quad (18)$$

$$\varepsilon_G(s) = Z(s) - G_K(s) \quad (19)$$

$$\varepsilon_R(s) = R(s) - R_K(s) \quad (20)$$

$$\beta = \sigma_{\varepsilon_G(s)}^2 + \sigma_{\varepsilon_R(s)}^2 \quad (21)$$

Here,  $Z(s)$  is the true rainfall field at location s.  $G_K(s)$  is the Krigeed (mean field) estimate of  $Z(s)$  from the rain gauge data,  $R(s)$  is the radar rainfall estimate,  $R_K(s)$  is the Krigeed (mean field) estimate of  $R(s)$  using the radar values at the rain gauge location and  $M(s)$  is the merged estimate of  $Z(s)$ .

The conclusion was extracted from this paper that spatial detail of the final merged field is improved while maintaining the mean field measured by the gauges.

### **|Modified conditional merging**

The other study modified conditional method to improve rainfall estimation considering elevation was presented by Yoon and Bae (2013). In that study, authors using elevation as secondary data to improve rainfall spatial distribution. The input data from rainfall and elevation were standardized to remove scale and dimension. Then the standardized ordinary cokriging is used (Isaaks & Srivastava, 1989). The rescaled ordinary cokriging follow:

$$Z^{1*} = \sum_{i=1}^{N_C} \lambda_i Z_i^1 + \sum_{j=1}^{M_C} \lambda_j (Z_j^2 - m_2 + m_1) \quad (22)$$

Here,  $Z^{1*}$  is a Kriged primary variable at all grid points that are either standardized kriged gauge rainfall or kriged radar rainfall. The  $Z_i^1$  is the standardized value of the primary variable at gridded location s obtained from collated rainfall locations  $i=1, \dots, N_C$ .  $Z_j^2$  is the standardized value of the secondary variable at locations obtained from collated rainfall locations  $i=1, \dots, M_C$ .  $\lambda_i, \lambda_j$  are cokriging weight.  $N_C, M_C$  are the numbers of primary and secondary data at collocated range, respectively. The  $m_1, m_2$  are estimated means of primary and secondary values computed from collocated pints. The ordinary cokriging possesses two unbiased constraints, while the rescaled ordinary cokriging has a single constrain that requires the sum of weights of all primary and secondary variables equal 1:

$$\sum_{i=1}^{N_C} \lambda_i + \sum_{j=1}^{M_C} \lambda_j = 1 \quad (23)$$

*The conclusion was extracted that method was compared to mean field bias method and ordinary conditional merging. Results show that it is a lesser tendency over smooths, and optimal mean areal rainfall like the value obtained using gauges.*

## **2. Discussion**

(1) We understand the problem in our proposed methodology. It needs to find the other solution such as the modified conditional merging method. But note that in this method, rainfall and elevation is determined that has a correlation. In our case, we have not determined the correlation of spatial drought to other climate variables. The NDI was computed from precipitation, runoff, soil moisture by principle component analysis. We also have satellite-based data precipitation,

runoff, soil moisture, but they are various to these inputs of NDI. Therefore, it is hard to find the correlation. In the next step, we should find the way to find the relationship between NDI and satellite-based data. Based on that we could consider using modified conditional merging method.

(2) ANN is robust to extract information from observation data. The results are obviously based on the training process. It depends on the quantity and quality of input data, also algorithm. We should think about how to enrich our data. Because if we only used data for specific extreme event in September 2015. It could be an insufficient. We proposed should choose several drought events from 2000 to 2016. We chose this period because the satellite images are almost available in this duration.

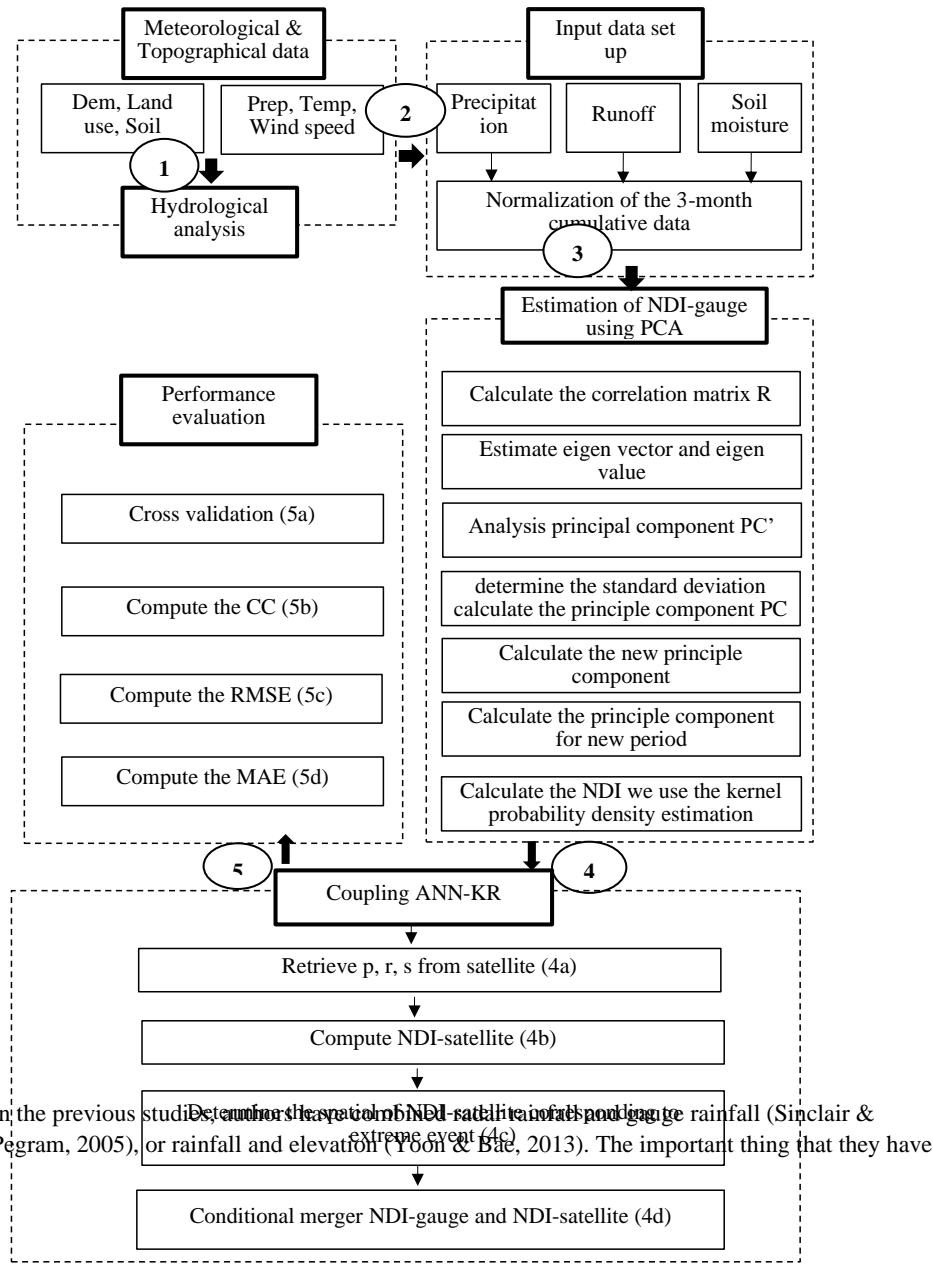
## References

- Cressie, N. (1993). *Statistics for spatial data*: John Wiley & Sons.
- Frajewski, W. F. (1987). Cokriging Radar-Rainfall and Rain Gage Data. In *1987, Rainfall Fields: Estimation, Analysis, and Prediction* (pp. 9571-9580).
- Isaaks, E. H., & Srivastava, R. M. (1989). *An Introduction to Applied Geostatistics*: Oxford University Press.
- Sinclair, S., & Pegram, G. (2005). Combining radar and rain gauge rainfall estimates using conditional merging. *6*(1), 19-22. doi:10.1002/asl.85
- Yoon, S.-S., & Bae, D.-H. (2013). Optimal Rainfall Estimation by Considering Elevation in the Han River Basin, South Korea. *Journal of Applied Meteorology and Climatology*, *52*(4), 802-818. doi:10.1175/JAMC-D-11-0147.1 %J Journal of Applied Meteorology and Climatology

## WEEK 34

A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network

### 3. Find the way to combine ANN and modified conditional merging



*Fig 104. Framework of spatial extreme drought coverage assessment*

a clear relationship. In our case, we have not found the relationship between NDI and other climate variables. Because our study using multiple variable drought index. NDI was computed from precipitation, runoff, soil moisture. The principle component analysis was used to reduce dimension in NDI. Therefore, it is hard to find the correlation between multiple drought index NDI and single climate variable. The results using ANN to predict NDI and compare to NDI-gauge shows very low correlation in the previous calculation (Fig 2).

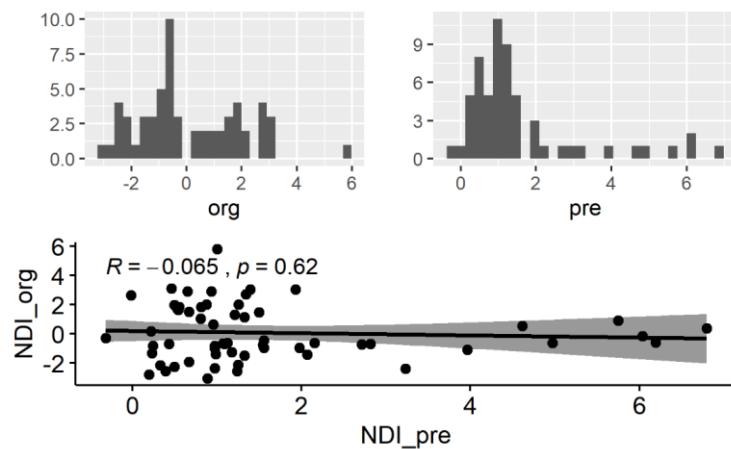


Fig 105. Histogram of original NDI, predicted NDI, and correlation.

To determine using the same type of primary and secondary data gives the high correlation, we had computed the SPI for South Korea based on remote sensing. The SPI-satellite, and SPI-gauge were computed Pearson correlation. Results show that they have a higher correlation ( $R= 0.85$ ) at the confidential level 99% (P value = 2.2e-16) in Fig 3.

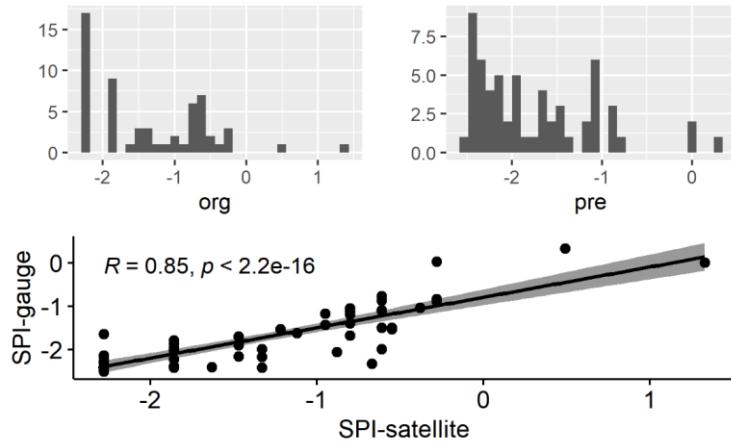


Fig 106. Histogram of original SPI-gauge, and predicted SPI-satellite, and correlation.

#### 4. Discussion

We determine the correlation of primary data and auxiliary data should be the same or obviously indeed relation. For example, the correlation should be considered between rain to rain, rain to elevation, SPI to SPI. We proposed indirectly using satellite-based data support spatial drought coverage assessment. The satellite-based data should be transformed to sample to gauge data type. In our case, satellite based should be translated to NDI-satellite to be an adopt for NDI-gauge.

#### References

- Sinclair, S., & Pegram, G. (2005). Combining radar and rain gauge rainfall estimates using conditional merging. 6(1), 19-22. doi:10.1002/asl.85  
 Yoon, S.-S., & Bae, D.-H. (2013). Optimal Rainfall Estimation by Considering Elevation in the Han River Basin, South Korea. *Journal of Applied Meteorology and Climatology*, 52(4), 802-818. doi:10.1175/JAMC-D-11-0147.1 %J Journal of Applied Meteorology and Climatology

## WEEK 35

A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network

### 5. Find the way to combine ANN and modified conditional merging

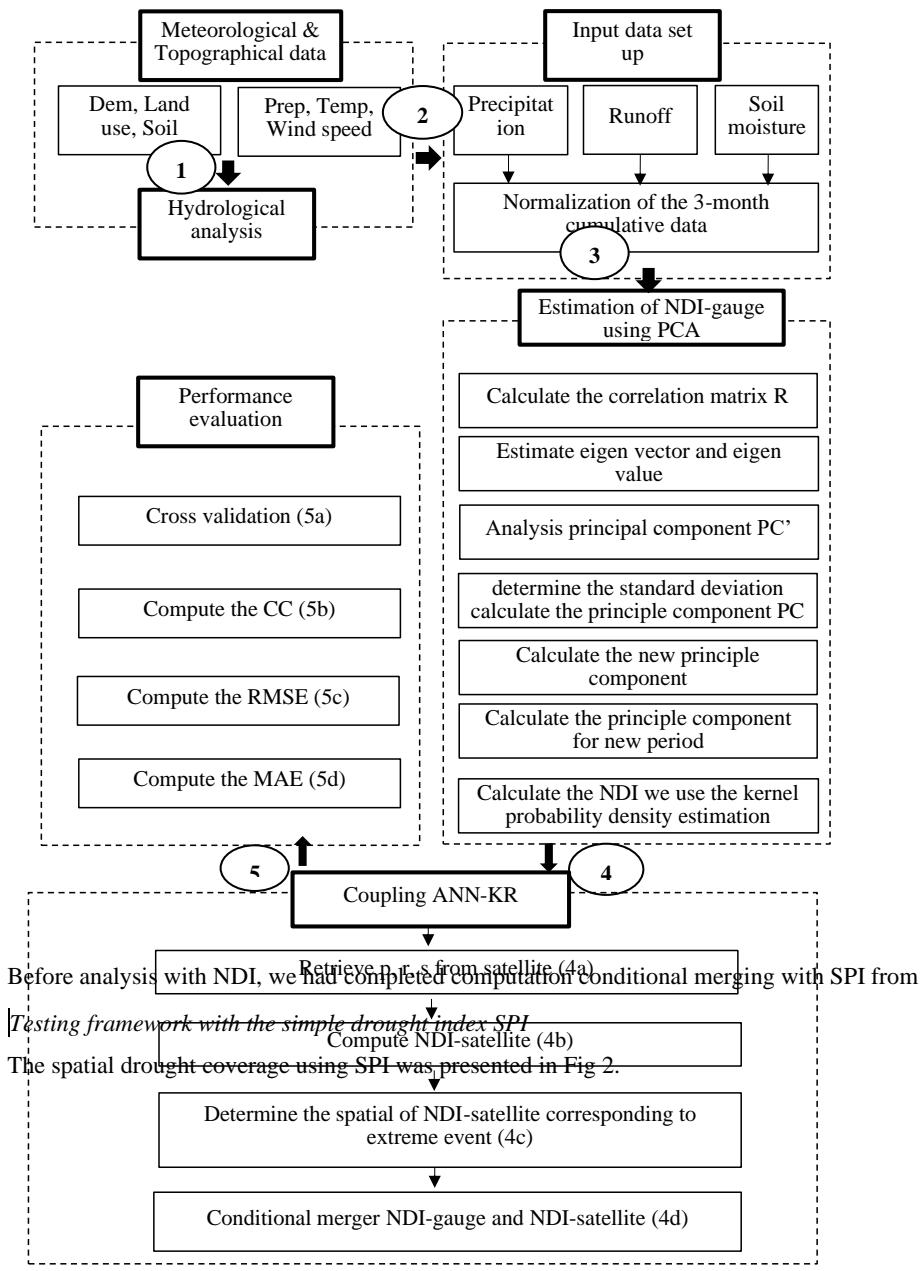


Fig 107. Framework of spatial extreme drought coverage assessment

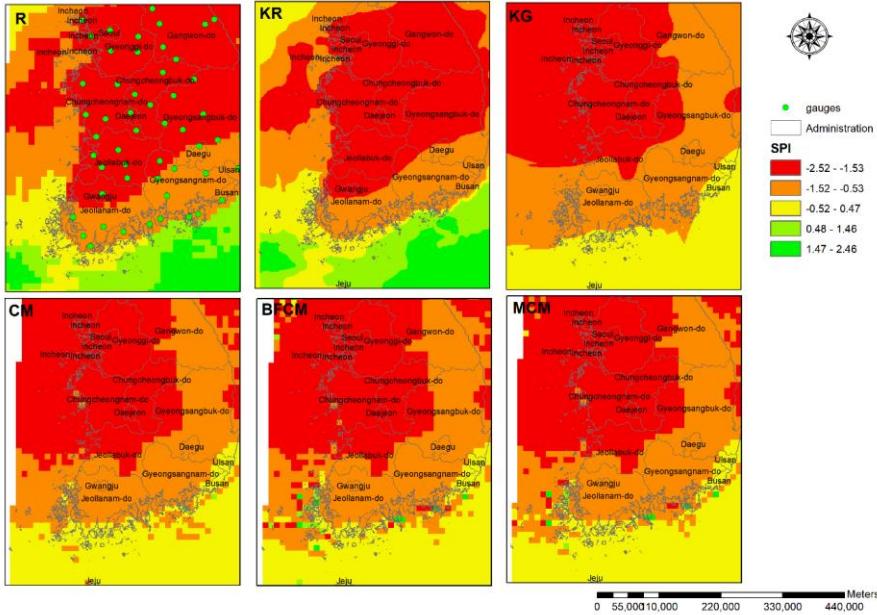


Fig 108. The SPI computed from satellite (R), kriging model from satellite points (KR), Kriging gauge model (KG), conditional merging (CM), Bias field conditional merging (BFCM), and mean conditional merging (MCM).

Results show that in the south of South Korea have high value SPI (green), while using kriging data from gauges it is lower (yellow). The combine of two type data R and KG give the flexible results of CM, BFCM, and MCM. To figure out the relationship between these results, the carter plots (Fig 3), and correlation matrix (Fig 4) was computed. We chose location at 59 ASOS for comparison because at these points we have the gauge-based data values. However, the correlation of G with KG, and other methods is not much difference (0.99-1.00). Therefore, it is hard to judge a model based on correlation in our case.

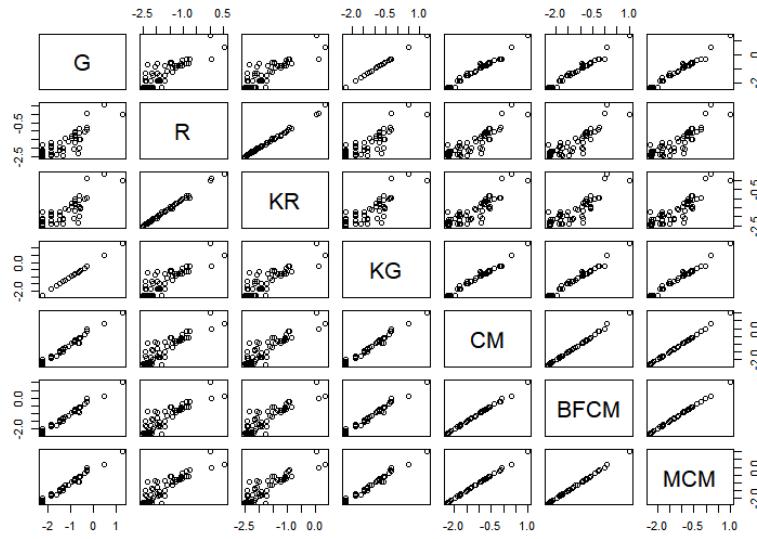


Fig 109. Scatter plot present relationship of G, R, KR, KG, CM, BFCM, and MCM.

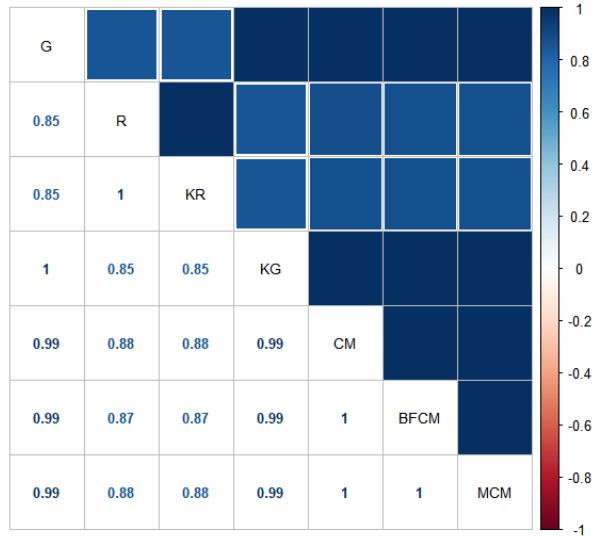
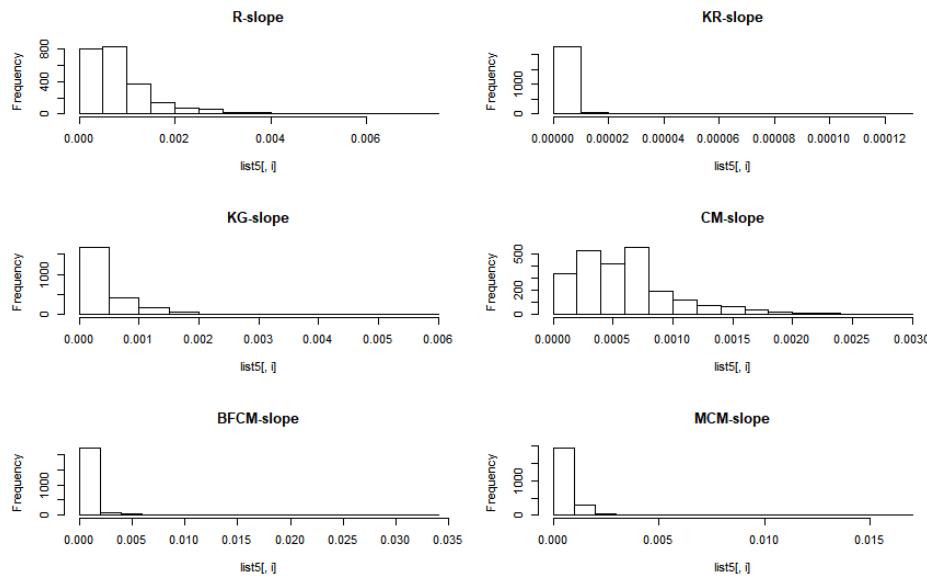


Fig 110. Matrix correlation of various interpolation methods.

### **|Improve model estimation**

Because using locations at 59 ASOS is difficult to estimate models, we propose using all the points at the center of grids to compare together. In addition, we use the computed surface slope of various models. We assume “slope” presents for variance of spatial distribution. The historical graph of various models was presented in Fig 5.



*Fig 111. Historical graph of all grid points from various models.*

Based on historical graph and correlation matrices of R-slope to others, BFCM and MCM is the most suitable model for spatial drought coverage.

### **6. Discussion**

(1) We proposed to limit our study with 2 assumptions. First, satellite (R) is a better spatial distribution compare to gauge (G) data. Value of specific gauge of G is better than R. Therefore, the predicted model (M) is determined that is the closest mean value of G, and having variance like to R.

(2) Based on this assumption, we determine a “surface” M has mean value is the closest to G, and slope is the closest to R. Obviously, these are two parameters should be considered: mean value and slope. The sequential method that solves a problem step by step. First fit the mean values follow to G, then fit the slope (SL) follow R. Similarly, The R can be fitted at the first, then the value V will be fitted later.

(3) The other approach that stimulates fit R and V at the same time. We determine the model M surface base on joint probability of V and G. After, training model with ANN, we determine the value of each grid point where are satisfied two conditions above.

(4) We still get stuck on finding how to use ANN in our study. We would like to keep finding the reasonable methods for our study. Therefore, in the next steps the main objective can apply ANN. The computation NDI from satellite data is postponed.

Professor comments:

Your direction is various with the idea that I thought at begin  
When you have time look back the  
Thinking why?  
But you can continue with your ideas in 2 weeks

## **WEEK 36**

**A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network**

### **7. Find the way to combine ANN and modified conditional merging**

There are 2 phases on our processing. First, we will find the correlation of gauges and satellite data. Second, we are using ANN to extract information from satellite-data. This information will be used to support for gauge-based data analysis. Normally, the primary and secondary data should have a relationship, or as the same type. For instance, radar rainfall is used to support gauges rainfall. Satellite-based soil moisture is used to support for observational gauged-based soil moisture. It should be precipitation - precipitation, soil moisture -soil moisture. In our study, we proposed to compute NDI based on satellite data because of these reasons.

NDI is computed from precipitation, runoff, soil moisture. These data we retrieved from various models. Precipitation was collected from the Famine Early Warning Systems Network (FEWS NET) Land Data Assimilation System (FLDAS) products. The FLDAS include a crop water balance model , Africa-specific daily rainfall from NOAA Climate Prediction center, and CHIRPS, a quasi-global rainfall dataset designed for seasonal drought monitoring and trend analysis(Funk et al., 2015). Runoff was also collected from FLDAS. Soil moisture is retrieved from the NASA Global Land Data Assimilation System (GLDAS). The goal of GLDAS is to generate optimal fields of land surface states and fluxes, by ingesting satellite- and ground-based observational data products. It is using an advanced land surface modelling and data assimilation techniques (Rodell et al., 2004). We subset these data for South Korea with monthly temporal scale. The spatial resolution of them are range from 0.10 to 0.25 degrees. The retrieved period from January 2000 to December 2016. In Fig 1 present map of soil moisture for South Korea in September 2015. We convert the raster to centroid points, then computed NDI by using principle component analysis (PCA). The question are given here that how we make sure that is existing relationship between

NDI-based on gauges and NDI-based on satellite. There are several sources lead to the bias of them. For testing, we analysis with SPI. The SPI was chosen for testing because it is simpler than NDI. Computation of SPI is based only precipitation. We improved the comparison not only spatial distribution but considered the sequence of SPI.

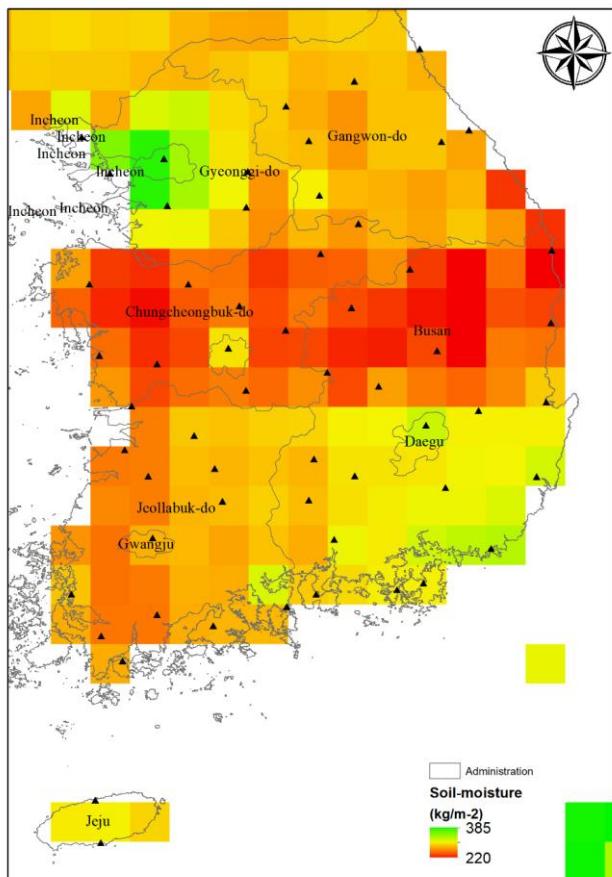
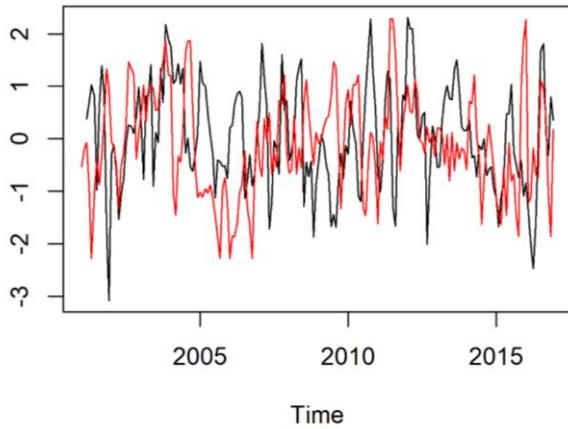


Fig 112. Soil moisture of South Korea in September 2015.

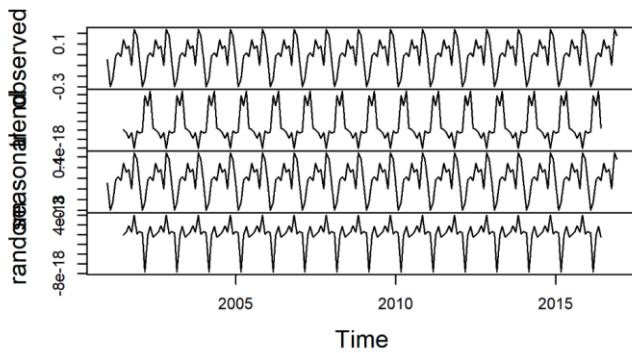
We chose Sokcho station for comparing. The time series SPI of Sokcho is presented in Fig 2.



*Fig 113. Compare SPI computed from satellite (red) and gauges (black).*

The sequence data was decomposed to random, seasonal, trend, observed part for deeply comparison. Fig 3 presents these components form SPI time series at Sokcho station.

#### Decomposition of additive time series



*Fig 114. The component of SPI computed from gauges.*

Similarly, we also analysis components with SPI computed from satellite (Fig 4)

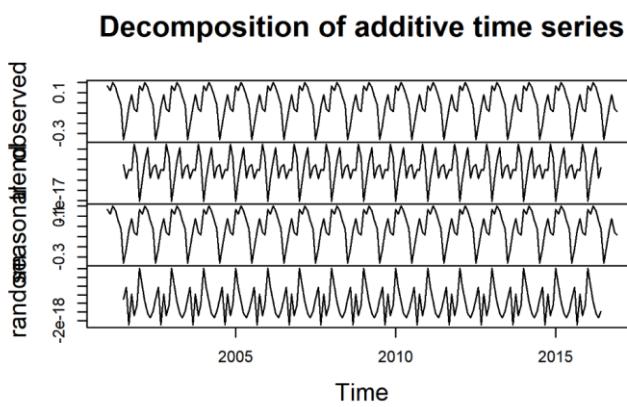


Fig 115. The component of SPI computed from satellite.

Unfortunately, we still can not find the relationship of sequence of them. Therefore, the ideas using both temporal and spatial characteristic of gauged data and satellite to find the correlation have not been obtained (Fig 5).

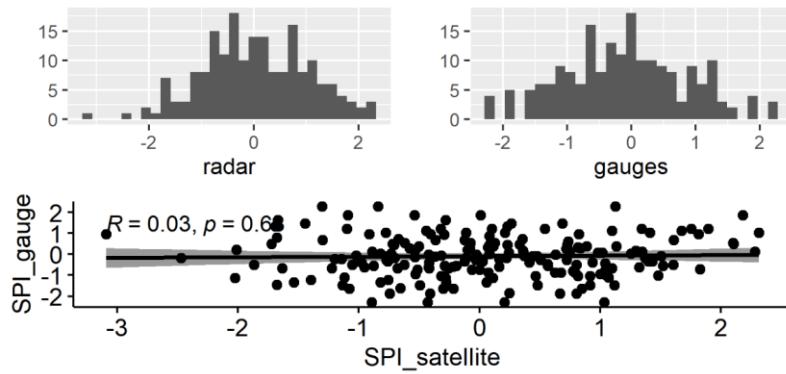


Fig 116. Correlation of SPI computed by gauges and satellite.

#### 8. Look back the initial idea

The original idea that combination of well spatial distribution of satellite-based data and more accuracy of gauge-based to improve spatial drought distribution. As the results of testing about, there is non-linearity correlation between of satellite-based data and gauge-based data, we should use ANN in this case to find the relationship of them. Then we will use a geostatistical model to

combine them to improve the spatial prediction. Conditional merging and Cokring model are among of the most suitable chosen.

## 9. Discussion

We are understood the challenge of our study is figure out the relation of satellite data and gauge-based data which is hard to be obtained by traditional methods. In the next steps, we propose a systematic review application of ANN in spatial analysis to find the solution. The rest of work will be temporal purposed.

## References

- Funk, C., Shukla, S., Hoell, A., & Livneh, B. J. B. o. t. A. M. S. (2015). Assessing the contributions of East African and west Pacific warming to the 2014 boreal spring East African drought. *96*(12), S77-S82.
- Rodell, M., Houser, P., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C.-J., . . . Bosilovich, M. J. B. o. t. A. M. S. (2004). The global land data assimilation system. *85*(3), 381-394.

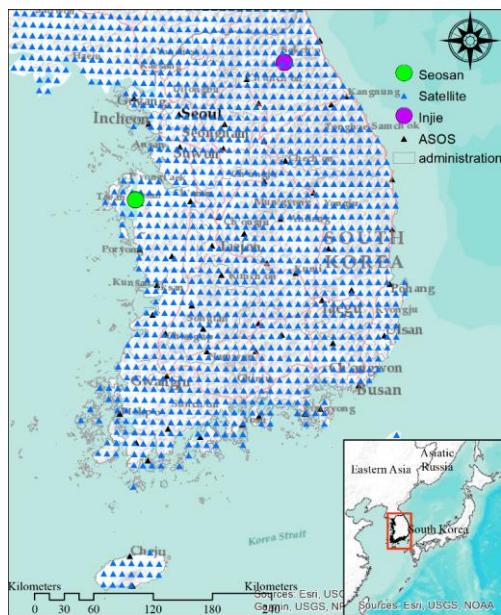
## WEEK 37

### A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network

After examining several satellite data and try to find the relationship to NDI-gauges, we extracted some initial conclusions.

#### 10. The computation NDI from satellite data

We retrieved precipitation, runoff, soil moisture from various satellite-based data. The data were structured to match with data from the gauges. The data were built increasingly from 59 stations to 1361 stations within the study area of 100,210 km<sup>2</sup>. The density of number stations was improved 23 times, from 1 station/1,698 km<sup>2</sup> to 1 station/73 km<sup>2</sup>. We chose Inje station and Seosan station for testing the correlation of precipitation, runoff, and soil moisture. These stations have the highest drought severity and a drought duration of 36 years (1981-2016).



#### |Compare input data of NDI

The comparative period is from 2000-2016 where the satellite data are available.

Fig 117. Study area, gauges data, satellite data and testing sample.

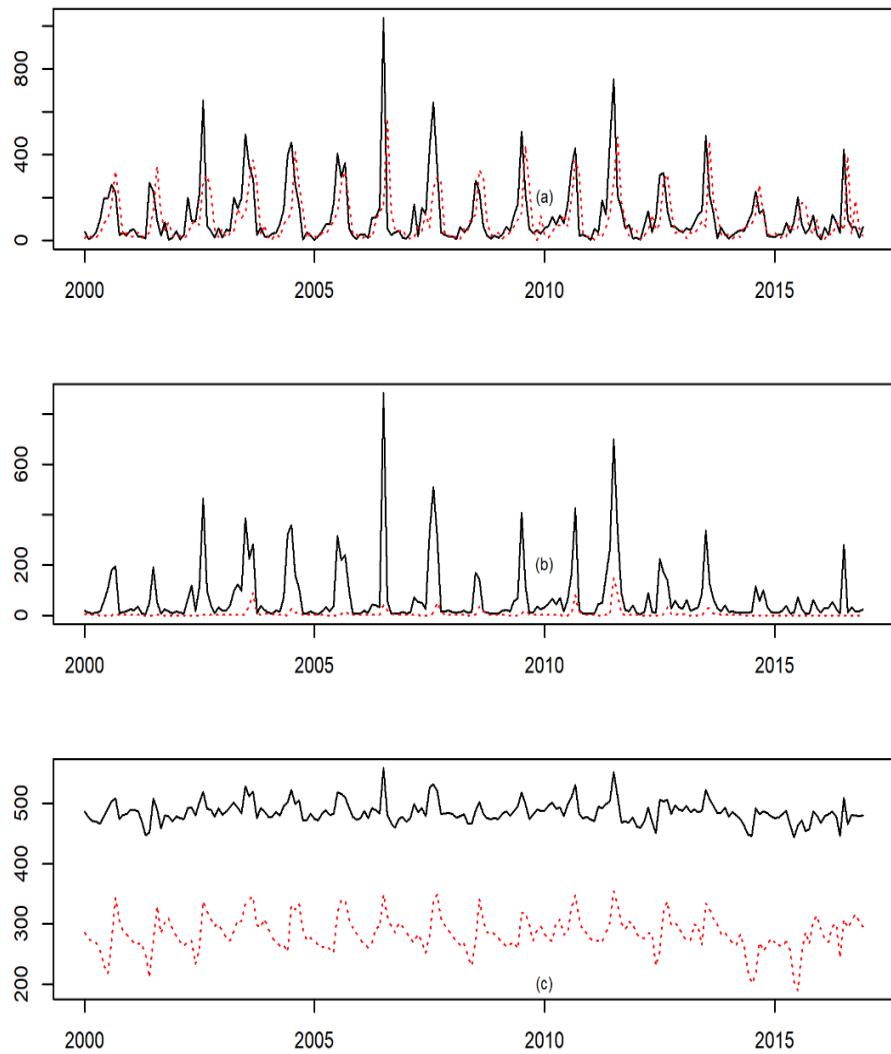


Fig 118. Compare rainfall (a), runoff (b), soil moisture (c) from gauge (black) and satellite (red) data for Injie station.

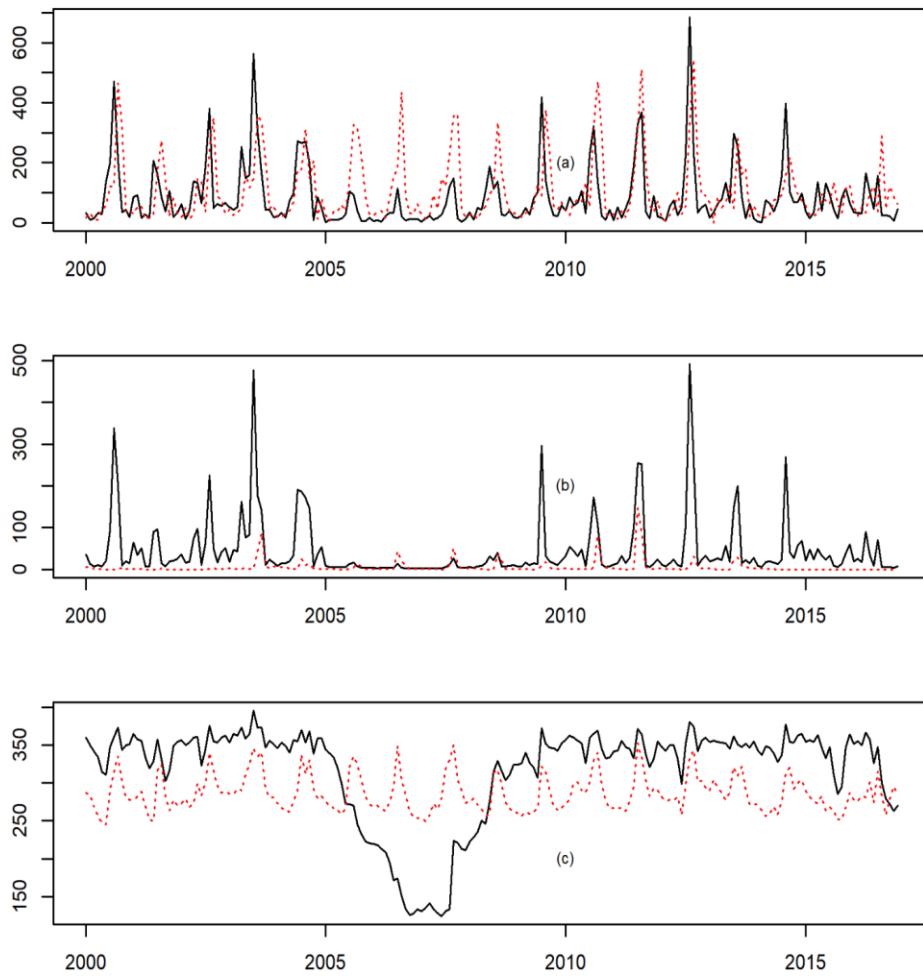


Fig 119. Compare rainfall (a), runoff (b), soil moisture (c) from gauge (black) and satellite (red) data for Seosan station.

In the Fig 2 and Fig 3 presents the comparison at Inje and Seosan station. Results shows only precipitation has the low correlation, runoff and soil moisture have a strong bias.

#### |Compare NDI computed from gauges and satellite

Fig 4 presents the result of NDI at Seosan station, which was computed by gauges and satellite.

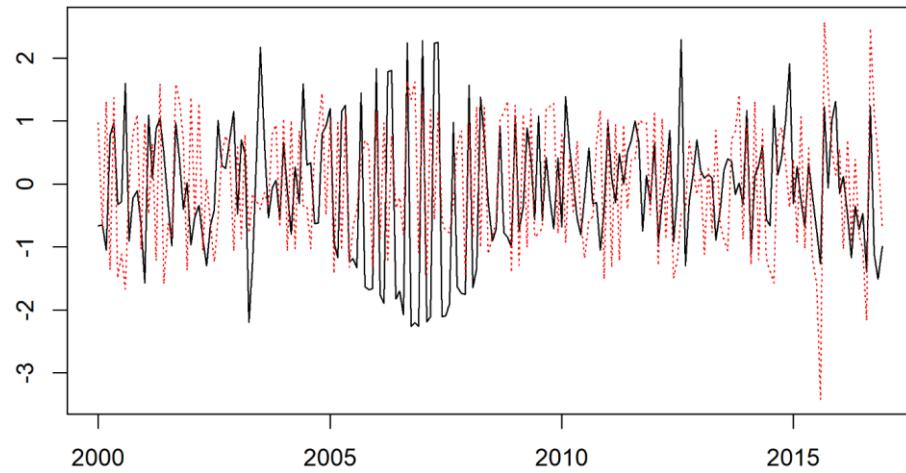
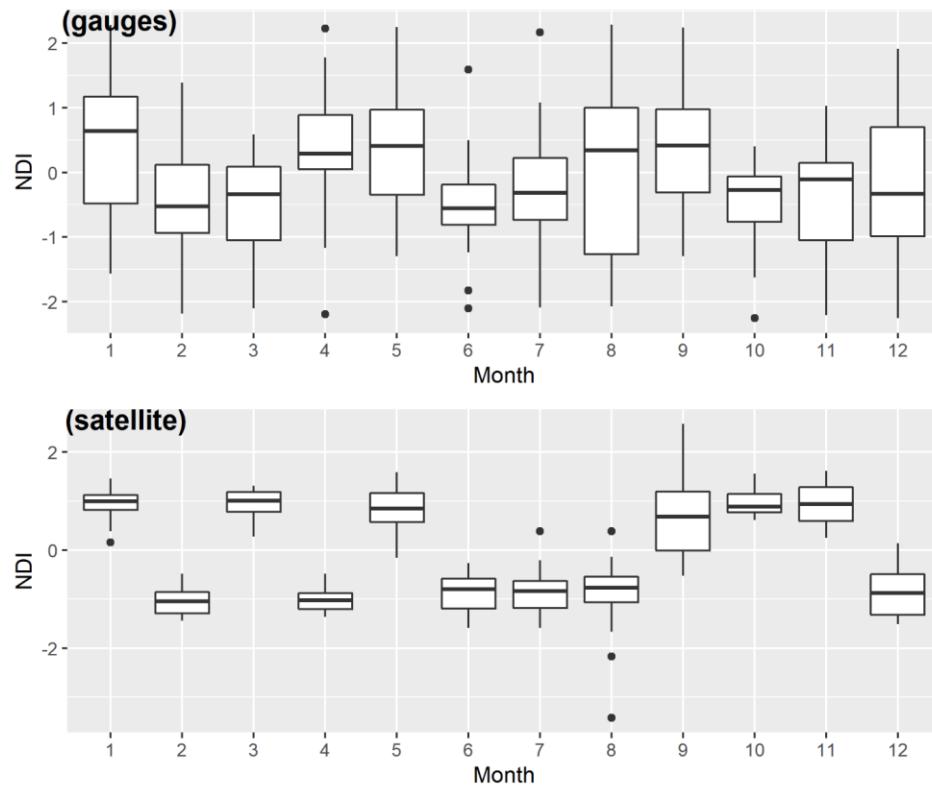


Fig 120. NDI at Seosan station computed from gauges (black), and satellite (red).

The monthly statistic shows the NDI-gauge has range is higher than NDI-sat in almost of years (Fig 5). The NDI - satellite does not show drought in the January from 2000-2016. All NDI is higher than 0. Otherwise, in Decembers, it shows this month almost in drought condition. The NDI computed from gauges is more flexible. Both drought and un-drought condition can occur every month. It is closer to the real condition than satellite. Therefore, it is hard to find the correlation between of NDI computed from gauges (NDI-gauge) and NDI computed from satellite (NDI-satellite).



*Fig 121. Monthly statistic NDI computed from gauges and satellite.*

### 11. Discussion

(12) The concept of using satellite-data to support gauge-data in the improve spatial drought assessment is interesting. It is reasonable when we find the relationship of them. It should be considered as the constraint between the primary and secondary data. When this correlation is revealed, it is an advantage to integrated both types of data. The multiple drought index NDI and satellite variables have a little correlation. Even we computed NDI from satellite to find the correlation of them, we still have not obtained it. Therefore, this issue should be handled before

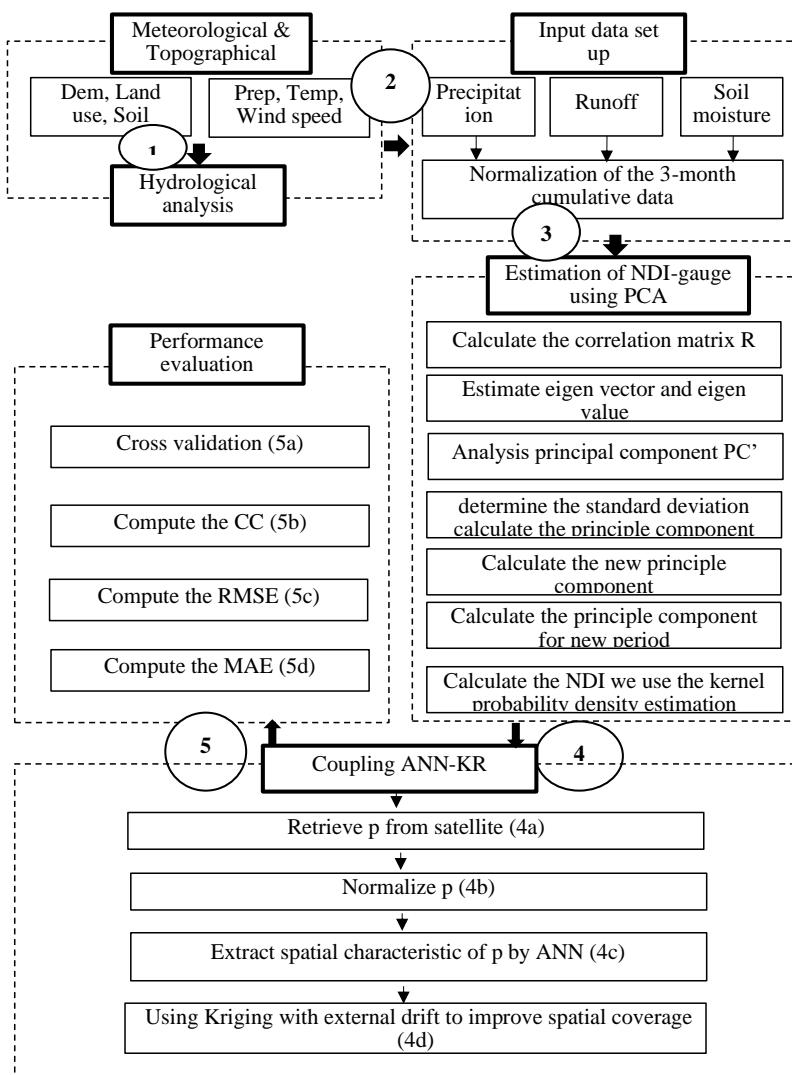
setting a new model. Neuron network model, Geostatistics model, or other models work well if they have the suitable data input. The subject and support object should have a relation.

(13) We definite the relation in spatial analysis as the autocorrelation between coordinate (x, y) and value (z). The auxiliary variables (m, n, k) support to improve spatial analysis when they have some ways to connect to z. It makes sense when we use satellite-precipitation and gauge precipitation to improve spatial precipitation. It is consistent to previous study. The previous study, we estimate extreme drought using NDI, copula and  $G_i^*$  statistic. We are going to research in deep and narrow issues. The aim of the current study is a proposal and method improve precipitation, one of the important inputs of NDI.

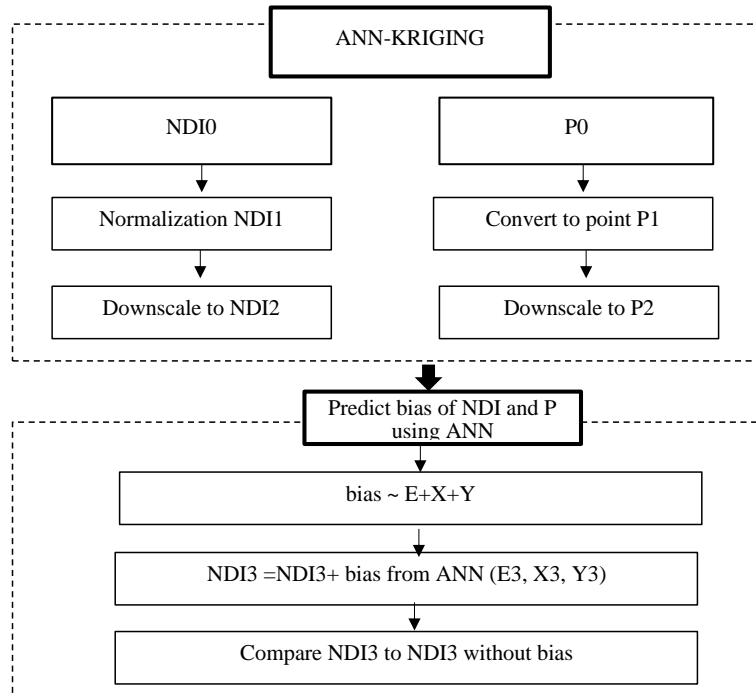
## WEEK 38

**A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network**

### 12. Set up framework



*Fig 122. Framework of spatial extreme drought coverage assessment*



We revised our methodology at step 4. In this step we use precipitation (p) as the auxiliary for support spatial distribution of NDI. The NDI is considered as “accuracy” source with limited spatial. While The p is considered as the “well” spatial coverage with less “accuracy”. Two source’s data above are normalized, fitted with the same scale. The bias of NDI and P is obtained from ANN model. It is used as the drift in final kriging external drift. Detail of step 4 is summarized in Fig 2.

The original NDI0 consist of 59 ASOS is normalized to remove dimension. Then it is downscaled to spatial resolution of 0.01 degrees, approximate 1 km. We used Krige as a tool to change spatial resolution. The precipitation image (P0) is covert to centroid point (P1). It is also normalized and transform to the same spatial resolution (P2) as NDI2. The variogram parameters are optimal to get the best fit for both NDI, P. In the next step, we add coordinate (X, Y) and topographic (E) as predictors of the model. Then, we predict the scalar value bias of NDI and P. This value will be used to correct spatial distribution of kriging NDI0.

*Fig 123. Predict bias of NDI and precipitation (P) based on elevation (E), coordinate system (X, Y).*

## **Discussion**

- (1) The downscale to get the higher spatial resolution helps us overcome the limited data. We obtained the huge dataset for ANN model. However, the spatial resolution at 1 km is not suitable. The computational time is increased. Several points have the same values when they are located at the same grid points. Therefore, we will try training data with a spatial resolution at 0.1 degrees in the next times.
- (2) The combine ANN and Kriging model based on a simple concept. The predicted value at unobserved point will be reduced error by removing the bias of them with the “accurate” auxiliary variables. It should be tested, revised before making a conclusion.

## WEEK 39

**A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network**

### 13. Overview spatial interpolation modes

Improve interpolation from sparse data is the main objective of our study. Therefore, we review some popular interpolation methods. Based on the correlation of drought and precipitation, we reviewed interpolation models using in precipitation instead of drought. Because it is easier assessing interpolation results from precipitation gauges than drought. The observed precipitation is directly measured, but drought index is indirectly computed. Three interpolation models were examined: Thiessen Polygon (TP), Inverse Distance Weight (IDW), Ordinary Kriging (OK), Kriging with an external drift (KED).

#### *|Thiessen Polygon*

This method is a simple that every ungauged location is allocated a value of closest gauge observation (H.Thiessen, 1911). The Study area is divided into the several polygons. Each polygon having a single gauge observation that presents for the entire area covered by that polygon. The main shortcoming of TP that it does not incorporate information about the neighboring gauge observation. Thus, results in abrupt jumps or discontinuities.

#### *|Inverse Distance Weight*

IDW assigns weights to neighboring observed values based on the distance to the interpolation location, and the interpolated value is the weighted average of the observations (Ahrens 2006). Measured values nearer to the predicted location have more influence on the predicted value than those farther away. The weight is computed by:

$$\lambda_i = \frac{1}{\sum_{i=1}^n \frac{1}{|D_i|^d}}, d > 0 \quad (24)$$

Where  $D_i$  is the distance between sample and unsampled points and  $d$  is the geometric form of the weight. Power  $d$  is usually set to 2 as inverse square distances. An optimal power value can be obtained by minimizing the RMSE.

#### *|Ordinary Kriging*

Different to two methods above, OK is a geostatistical method that uses a semivariogram to characterize spatial dependence. The semivariogram is computed as half of the square difference between paired values to the distance by which they are separated

$$\gamma(h) = \frac{1}{2N(h)} \sum_{\substack{(i,j) | d_{ij}=h \\ - Z_i)^2}} (Z_j - Z_i)^2 \quad (25)$$

Where  $2N(h)$  is number pair of data locations at distance  $h$ . Usually, the average squared distance for all pairs separated by a range of distances are obtained and plotted against the separation distance. A theoretical model is then fitted to the experimental semivariogram and its coefficients used for Kriging. Seven theoretical models exist, namely logarithmic, exponential, Gaussian, rational quadratic, spherical, penta-spherical, and power model.

#### *Kriging with an external drift*

Kriging with external drifts allows the incorporation of one or more additional variables that are used as background information for the interpolation of the primary variable. The basic assumption of KED is that the expected value of the estimated variable  $Z(u)$  has a relationship with an additional variable  $Y(u)$

$$E[Z(u)|Y(u)] = a + b \cdot Y(u) \quad (26)$$

A considerable problem when applying kriging with an external drift for merging of the station and satellite data is the frequent occurrence of numerical instabilities in the kriging system. In the study interpolated rainfall considering elevation, Goovaerts (2000) show that larger prediction errors are obtained for the two algorithms (inverse square distance, Thiessen polygon) that ignore both the elevation and rainfall records at surrounding stations. The three multivariate geostatistical algorithms outperform other interpolators, in particular the linear regression, which stresses the importance of accounting for spatially dependent rainfall observations in addition to the collocated elevation.

#### **14. Clarify the framework**

The main objective of the framework is combining satellite and gauged information to improve spatial drought distribution. The merging ( $NDI_F$ ) is the function of the satellite precipitation ( $P_S$ ) and NDI value ( $NDI_0$ ).

$$NDI_F = P_S + f(NDI_0 - P_S) \quad (27)$$

The  $f(NDI_0 - P_S)$  represents biases to be corrected in the original satellite precipitation data and it usually solve by certain mathematical optimization methods.

In this study, we propose using three steps satellite-gauge precipitation merging (Chao et al., 2018).

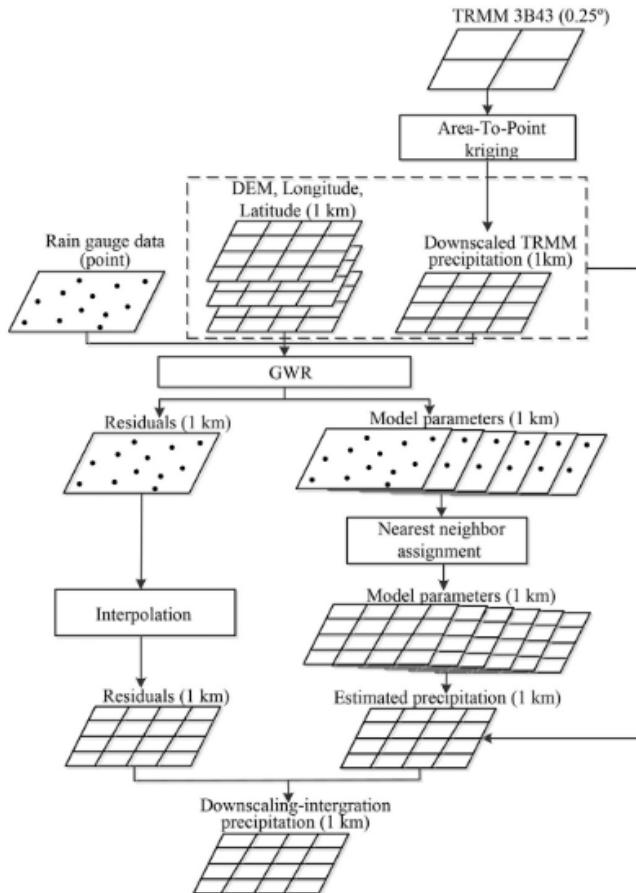


Fig 124. The flow chart of downscaling-integration.

We modified the framework by replacing ground-based gauges precipitation by ground-based NDI gauges. First, we used the bilinear interpolation to downscale coarser satellite precipitation data

$(P_M)$  from 10 km to finer spatial resolution at 1 km. Then applied the Geography Weight Regress (GWR) model to estimate the values of  $f(NDI_0 - P_S)$ , bias of downscale precipitation. Final, we remove the derived bias from down scale satellite precipitation to produce the final downscale merged precipitation data using equation 4. The GWR method is suitable for analyzing the spatial relationship between precipitation and its influencing factors such as longitude, latitude, elevation, and so on (Chao et al., 2018; Lv & Zhou, 2016; Zhou et al., 2016). Remember that all NDI and precipitation are normalized to be scalar without dimensional units. The detail of data fusion and application each step is present in the Fig 2 below.

Rather than using an optimization method to solve  $f(NDI_0 - P_S)$ , we setup a simple GWR model

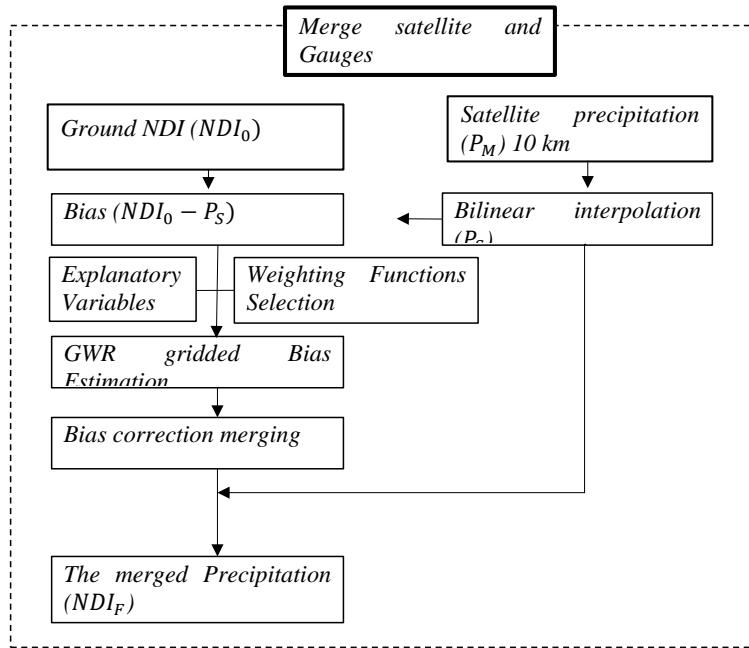


Fig 125. The flowchart of merging satellite and gauges data.

estimate  $f(NDI_0 - P_S)$  by Neural network with relating parameters such as elevation, slope, aspect. Assuming that biases in the satellite observations of precipitation are mainly depending on geographical, topographical, and meteorological characteristics. Following the typical GWR method (Brunsdon et al., 1996), the biases estimated by the GWR method can be written as:

$$f(NDI_0 - P_S) = b_i = a_{i0} + \sum_{k=1}^n a_{ik}x_{ik} + \varepsilon_i \quad (28)$$

Where  $b_i$  is the bias, difference between gauge NDI data and down-scale satellite precipitation data at location grid cell i;  $a_{i0}$  is the constant regression coefficient; n is the number of explanatory variables (elevation, aspect, slope).

The regression coefficients at a location of grid cells is estimated follow (Fotheringham et al., 1998):

$$\begin{aligned} a(i) \\ = (X^T w(i) X)^{-1} X^T w(i) b \end{aligned} \quad (29)$$

Where  $a(i)$  is the  $n \times 1$  vector of regression coefficients at location or grid cell i;  $w(i)$  is a diagonal matrix ( $m \times m$ ) of spatial weights obtained by the weighting functions quantifying the proximities of location I to its m neighbor gauge station; X is the variable matrix ( $m \times n$ ) of variable such as elevation, aspect, slope; and the bias vector ( $m \times 1$ ) is the difference between gauge data and downscaled satellite data.

## 15. Enhance data

Following comments, we retrieved satellite images from 2001-2016 to enhance data. Fig 3 presents the time series of monthly precipitation from 2000 to 2015. It is consistent with the previous results (Bae & Chang, 2019; Hong et al., 2016; Kwon et al., 2016). September 2000 was recorded as the flooding time in South Korea. The precipitation in the north of South Korea is higher than 300mm. Otherwise, September 2015 was determined as the most serve droughts. Because it has

the least precipitation in this time. More than a half of South Korea has precipitation under 130mm.

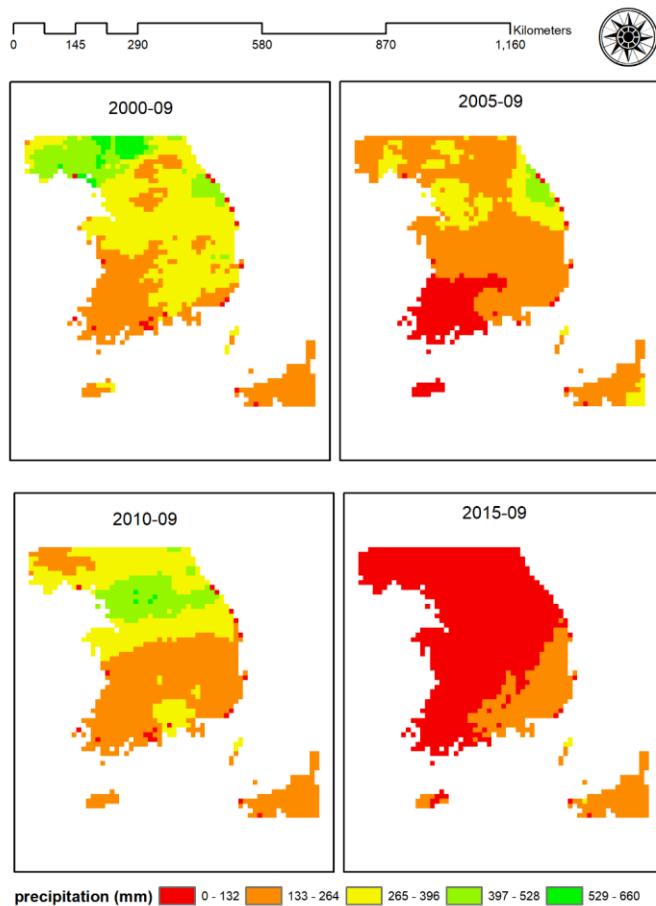


Fig 126. Monthly precipitation for South Korea at spatial resolution 10 km.

## 16. Discussion

- (1) After searching the solution for our study, we found several approaches to combine satellite and gauge data. The conditional merging is correct satellite and gauge remove bias based on fractions. The Geography weight regression using bias base on subtraction. They are

both based on as the same type of precipitation. In our case, we used two types NDI and precipitation with various dimensions. After removing dimension and scale by normalization we try to apply these methods above. However, we are still afraid of the relationship of scalar NDI and satellite precipitation. Does it look like to the satellite precipitation and ground based-precipitation gauges case? Should we find the way to prove it is the same condition?

ANN was not outperforming compared to geostatistical methods. It is worse than the radius basic function (Kampüs & Sciences, 2008). Estimate ANN as interpolation tool, Nevtipilova et al. (2014) figure out that ANN are not better results than using IDW and Kriging. Therefore, we should consider the other methods such as deep learning (Moraux et al., 2019), combination of ANN with other geostatistical methods.

(2) In the next step, we propose testing our framework with GWR methods. Then try to use ANN as a function instead of GWR.

## References

- Bae, S., & Chang, H. (2019). Urbanization and floods in the Seoul Metropolitan area of South Korea: What old maps tell us. *International Journal of Disaster Risk Reduction*, 37, 101186. doi:<https://doi.org/10.1016/j.ijdrr.2019.101186>
- Brunsdon, C., Fotheringham, A. S., & Charlton, M. E. (1996). Geographically weighted regression: a method for exploring spatial nonstationarity. *Geographical Analysis*, 28(4), 281-298. doi:[10.1111/j.1538-4632.1996.tb00936.x](https://doi.org/10.1111/j.1538-4632.1996.tb00936.x)
- Chao, L., Zhang, K., Li, Z., Zhu, Y., Wang, J., & Yu, Z. (2018). Geographically weighted regression based methods for merging satellite and gauge precipitation. *Journal of Hydrology*, 558, 275-289. doi:<https://doi.org/10.1016/j.jhydrol.2018.01.042>
- Fotheringham, A. S., Charlton, M. E., Brunsdon, C. J. E., & A, p. (1998). Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. 30(11), 1905-1927.
- Goovaerts, P. (2000). Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228(1), 113-129. doi:[https://doi.org/10.1016/S0022-1694\(00\)00144-X](https://doi.org/10.1016/S0022-1694(00)00144-X)
- H.Thiessen, A. (1911). Precipitation averages for large areas. *MONTHLY WEATHER REVIEW*, 39.
- Hong, I., Lee, J. H., & Cho, H. S. (2016). National drought management framework for drought preparedness in Korea (lessons from the 2014-2015 drought). *Water Policy*, 18(S2), 89-106. doi:[10.2166/wp.2016.015](https://doi.org/10.2166/wp.2016.015)

- Kampüs, K. J. I. A. o. P., Remote Sensing, & Sciences, S. I. (2008). Estimation of Unknown Height with Artificial Neural Network on Digital Terrain Model. 116-118.
- Kwon, H.-H., Lall, U., & Kim, S.-J. (2016). The unusual 2013-2015 drought in South Korea in the context of a multicentury precipitation record: Inferences from a nonstationary, multivariate, Bayesian copula model: The Unusual 2013-2015 Drought in South Korea. *Geophysical Research Letters*, 43(16), 8534-8544. doi:10.1002/2016GL070270
- Moraux, A., Dewitte, S., Cornelis, B., & Munteanu, A. (2019). Deep Learning for Precipitation Estimation from Satellite and Rain Gauges Measurements. 11(21), 2463.
- Nevtipilova, V., Pastwa, J., Boori, M. S., Vozenilek, V. J. J. o. G., & Geophysics. (2014). Testing artificial neural network (ANN) for spatial interpolation. 3(2), 01-09.

Professor comments:

-You are at mid process

-He only give us after read the final results

## WEEK 40

### A framework evaluates extreme drought coverage using natural drought index, Satellite based data and artificial neural network

#### 1. Parameters of satellite images

We examined extra parameter of satellite image to support for building model. It is expected the model “learn” more from these parameters. Precipitation quality index, gauge relative weighting, probability liquid precipitation, and random error were analyzed as follows.

##### *Precipitation quality index*

Precipitation quality index gives some guidance on when most should trust the Integrated Multi-satellite Retrieval for GPM (IMERG). At the monthly scale, the metric for random error, based on G. J. I. W. Huffman, DC, USA (2018) analysis of sampling error for a particular data sources. The general form of the relationship is:

$$\sigma_r^2 = \frac{\bar{r}^2}{r} \left( \frac{H}{p} - 1 \right) \quad (30)$$

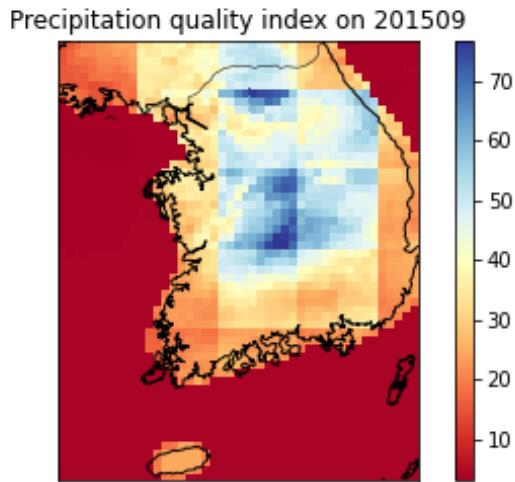
Where  $\sigma_r$  is random error,  $\bar{r}$  is the time-average of precipitation rate sample,  $N$  is the number of independent samples in  $\bar{r}$ ,  $H$  is the non-dimensional second moment of the probability distribution of the precipitation rates, and  $p$  is the frequency of all nonzero precipitation. Equation (1) is approximated as:

$$\sigma_r^2 \cong \frac{H}{I} \frac{(\bar{r} + S)}{N} [24 + \sqrt{\bar{r}}] \quad (31)$$

where  $\bar{r}$  and  $N$  are available for each grid box in the monthly estimate,  $I$  is a multiplicative constant expressing the fraction of  $N$  that is “independent”, and  $\frac{H}{I}$  and  $S$  are global constants that are approximated with validation data for each sensor type. Simple relationship as the inverted for  $N$ . All the constants are set gauge analysis, but the  $\bar{r}$  and  $\sigma_r^2$  used are the final satellite-gauge precipitation estimate and random error variance:

$$N \cong \left( \frac{H}{I} \right)_g \frac{\bar{r} + S_g}{\sigma_r^2} [24 + 49\sqrt{\bar{r}}] \quad (32)$$

Where  $N$  is defined as the equivalent number of gauges. Following Huffman (1997), the interpretation is that the approximate number of gauges required to produce the estimated random error, given estimated precipitation. The units are gauges per area. The current implementation area is  $2.5^\circ \times 2.5^\circ$  of latitude/longitude.



*Fig 127. Quality index for estimate monthly precipitation in South Korea.*

The center of South Korea is “high quality” as the number equivalent gauges from 60-70 gauges over 625 km<sup>2</sup>.

#### |Gauge relative weighting

Weighting of gauge precipitation relative to thematic-satellite precipitation. It was computed by the percentage of gauge precipitation relate to the multiple-satellite precipitation. In the Fig 2 shows the high weight in the center of South Korea over 70 percentage.

Gauge relative weighting on 201509

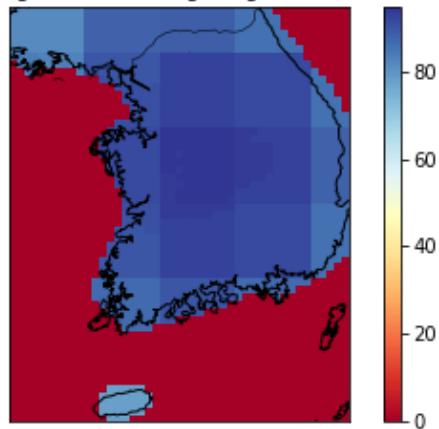


Fig 128. Weighting of gauge precipitation relative to the multi-satellite precipitation.

|Probability liquid precipitation

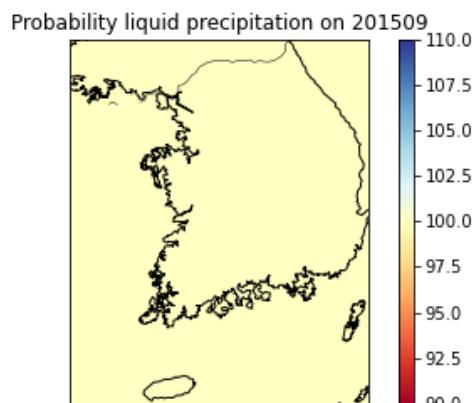


Fig 129. Probability of liquid precipitation phase (percent).

The probability liquid precipitation is computed in percentage of accumulation probability of liquid precipitation. It shows that the cumulative weight for a month is 100% (Fig3)

### |Random error

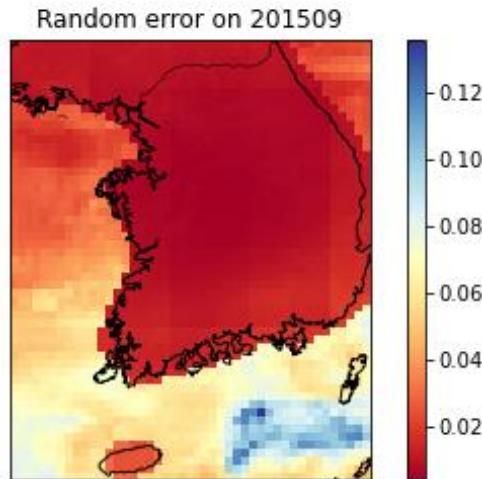


Fig 130. Random error (mm/h)

The random error of IMERG is computed follow (G. J. J. J. o. A. M. Huffman, 1997). The random errors contained in a finite set E of precipitation estimates result from both finite sampling and measurement-algorithm effects. The expected root-mean-square random error associated with the estimated average precipitation in E is shown in the equation (2) and explain above. It has a consistency of the other satellite's parameters. The center of South Korea has the lowest random error under 0.02mm/h.

## 2. Discussion

We examined other parameters of satellite because it could be considered as the input for our model. In addition, understand them could help us explain and discuss about our results. The quality of satellite images is not equal for whole the study area. Therefore, the results from framework has the highest confidence in the center of South Korea. Otherwise, the surrounding area is the lower confidence.

In the next step, we propose testing geographic weight regression, build the ANN with extra variables of precipitation quality index, gauge relative weighting, random error.

## **References**

- Huffman, G. J. I. W., DC, USA. (2018). IMERG Quality Index.
- Huffman, G. J. J. o. A. M. (1997). Estimates of root-mean-square random error for finite samples of estimated precipitation. *36*(9), 1191-1201.

## WEEK 41

### Weekly report detail

#### 1. Recent extreme drought trend in South Korea

**|Summarize framework for extreme drought assessment.**

The main objective of this study is determining extreme drought in South Korea by using Multiple Probability-Geostatistical methods. The spatial and temporal variability of drought in South Korea were investigated using 59 weather station data for 1981-2016. Various temporal scale 1, 3, 6-months of the Natural drought index was utilized to estimate drought. The natural drought index consists of meteorological (precipitation), hydrological (runoff), and agricultural drought (soil moisture). Precipitation was provided by the Korea Meteorological Administration. Runoff and soil moisture were extracted from the Variable Infiltration Capacity model. The continuous drought surface was modelled by using Kriging model. Maximum, minimum, mean, standard variables of the natural drought index were estimated for each administration of South Korea. The temporal-spatial trend, relationships of drought were analyzed by using the Mann-Kendall test, empirical mode decomposition (EMD) method, Moran's I test statistic, and Gi\* statistic. Results show that natural drought has complex trends and there is a strong spatial relationship in their variability. This variation is likely to increase drought risk cause of random, stochastic characteristic of drought phenomena.

**|Methodology**

The processing of estimate extreme drought consists of 4 steps (Fig 1). First step, Natural Drought Index (NDI) were computed as various temporal scales 1-, 3-, 6- months at 59 stations. The period was chosen from 1981 to 2016 where data are available. The second step is using Kriging model to build grid-cells of NDI at 1 degree ~ 10 km spatial scale with 1-month, 3-months, and 6-months temporal scale. In the third step, we estimate maximum, minimum, mean and variance of each administrative region. Total 200 administration units which have the area is larger than 25 km<sup>2</sup> were chosen for examination. In the final step, we used the Mann-Kendall, Empirical model decomposition, Moran's I statistic, Gi\* statistic to estimate tempo-spatial trend of extreme drought.

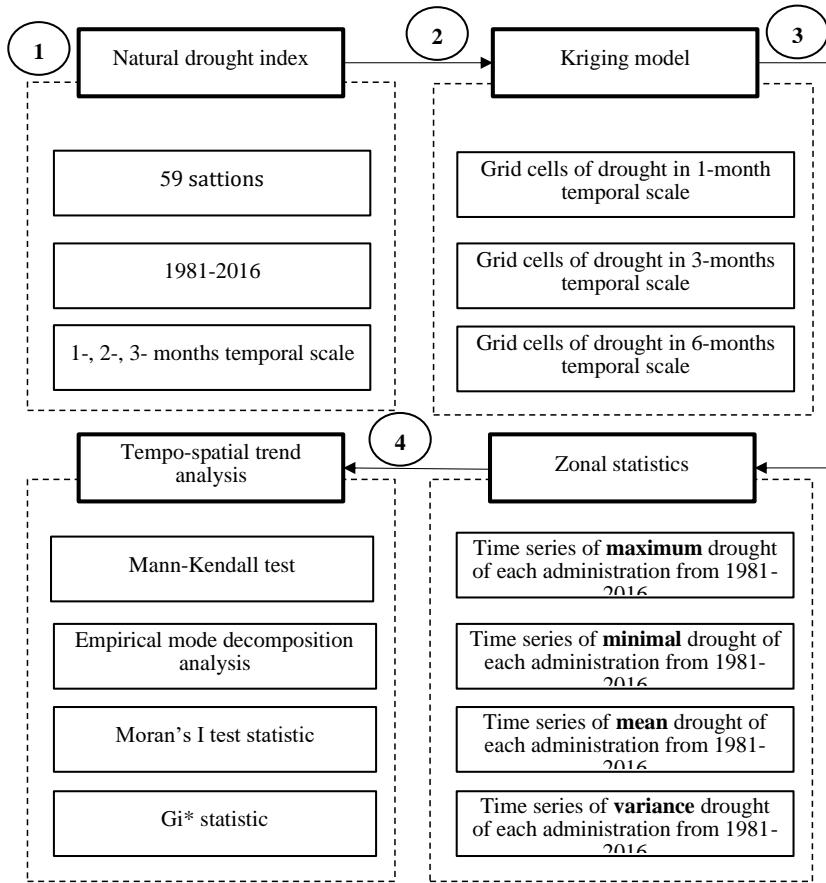


Fig 131. Flowchart for extreme drought trend assessment.

## 2. Geography Weight Regression

Spatial analysis included spatial heterogeneity and locality. Geography Weight Regression (GWR) is a type of local spatial prediction model that incorporates spatial homogenous into a regression model. GWR was explored in previous studies (Brunsdon et al., 1998; Chao et al., 2018; Chen et al., 2020). The GWR presents the spatially varying relationships are explored between the dependent and independent variables. Exploration commonly consists of mapping the resultant local regression coefficient estimates and associated (pseudo) t-values to determine evidence of non-stationarity. The basic form of the GW regression model is

$$y_i = \beta_{i0} + \sum_{k=1}^m \beta_{ik} x_{ik} + \epsilon_i \quad (33)$$

Where  $y_i$  is the dependent variable at the location,  $x_{ik}$  is the value of the  $k$ th independent variable at location  $i$ ;  $m$  is the number of independent variables;  $\beta_{ik}$  is the local regression coefficient for the  $k$ th independent variable at location  $i$ ; and  $\epsilon_i$  is the random error at  $i$ .

As data are geographically weighted, nearer observations have more influence in estimating the local set of regression coefficients than observations farther away. The model measures the inherent relationships around each regression point  $i$ , where each set of regression coefficients is estimated by a weighted least squares approach. The matrix expression for this estimation is:

$$\hat{\beta}_i = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) y \quad (34)$$

Where  $X$  is the matrix of the independent variables with a column 1<sup>st</sup> for the intercept;  $y$  is the dependent variable;  $\hat{\beta}_i = (\beta_{i0}, \dots, \beta_{im})^T$  is the vector of  $m + 1$  local regression coefficient; and

$W_i$  is the diagonal matrix denoting the geographical weighting of each observed data for regression point  $i$  at location  $(u_i, v_i)$ .

### 3. Discussion

We intend to parallel two studies: (i) analyze extreme drought trends, (ii) and improve spatial drought coverage. However, the first study should be complete before November 2020. Because the conference will be held on November 5<sup>th</sup> this year. Therefore, we propose taking a priority for the first study. The second study will be done after that.

### References

- Brunsdon, C., Fotheringham, S., & Charlton, M. J. J. o. t. R. S. S. D. (1998). Geographically weighted regression. *47(3)*, 431-443.
- Chao, L., Zhang, K., Li, Z., Zhu, Y., Wang, J., & Yu, Z. (2018). Geographically weighted regression based methods for merging satellite and gauge precipitation. *Journal of Hydrology*, *558*, 275-289. doi:<https://doi.org/10.1016/j.jhydrol.2018.01.042>

Chen, S., Xiong, L., Ma, Q., Kim, J.-S., Chen, J., & Xu, C.-Y. (2020). Improving daily spatial precipitation estimates by merging gauge observation with multiple satellite-based precipitation products based on the geographically weighted ridge regression method. *Journal of Hydrology*, 589, 125156. doi:<https://doi.org/10.1016/j.jhydrol.2020.125156>

Professoer comments:  
Go ahead

**WEEK 42**  
**Recent extreme drought trend in South Korea**

1. Explore data

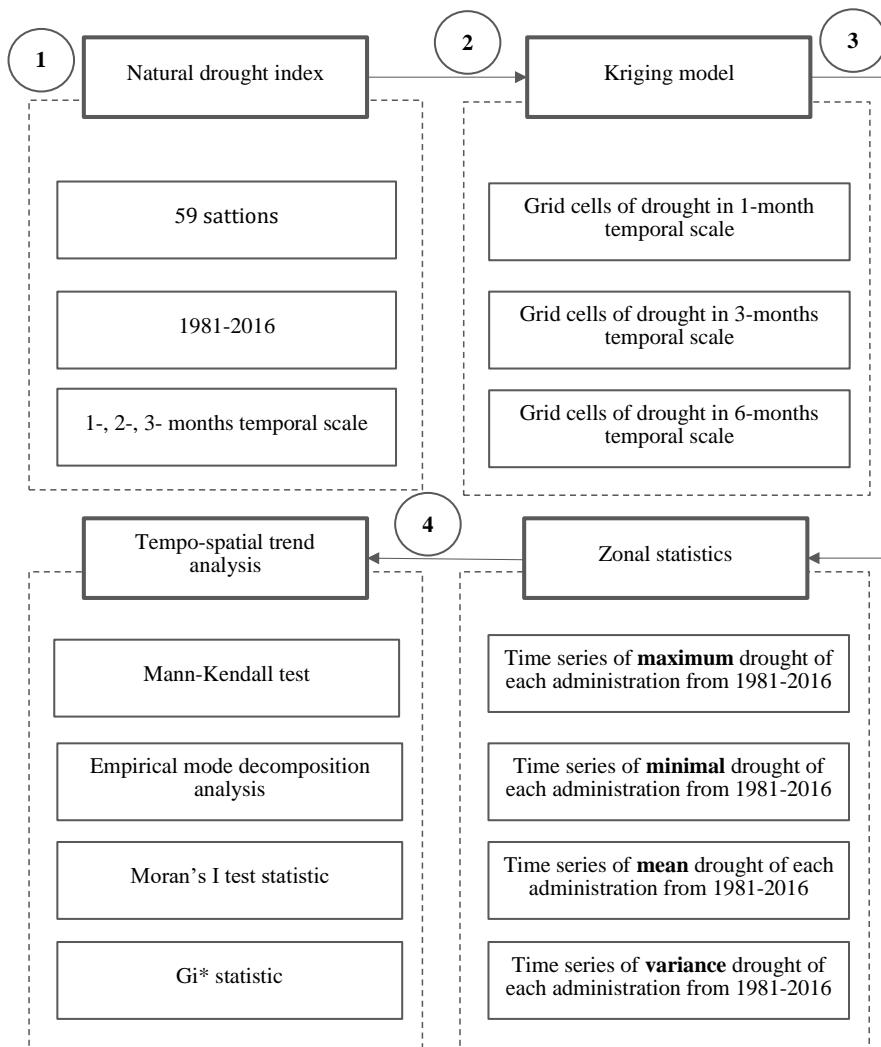


Fig 132. Flowchart for extreme drought trend assessment.

### |Spatial data explore

The input data consists of NDI at 59 ASOS locations were explored to figure out the general spatial characteristic. Distance, density, normality, random statistics were examined (Fig 2).

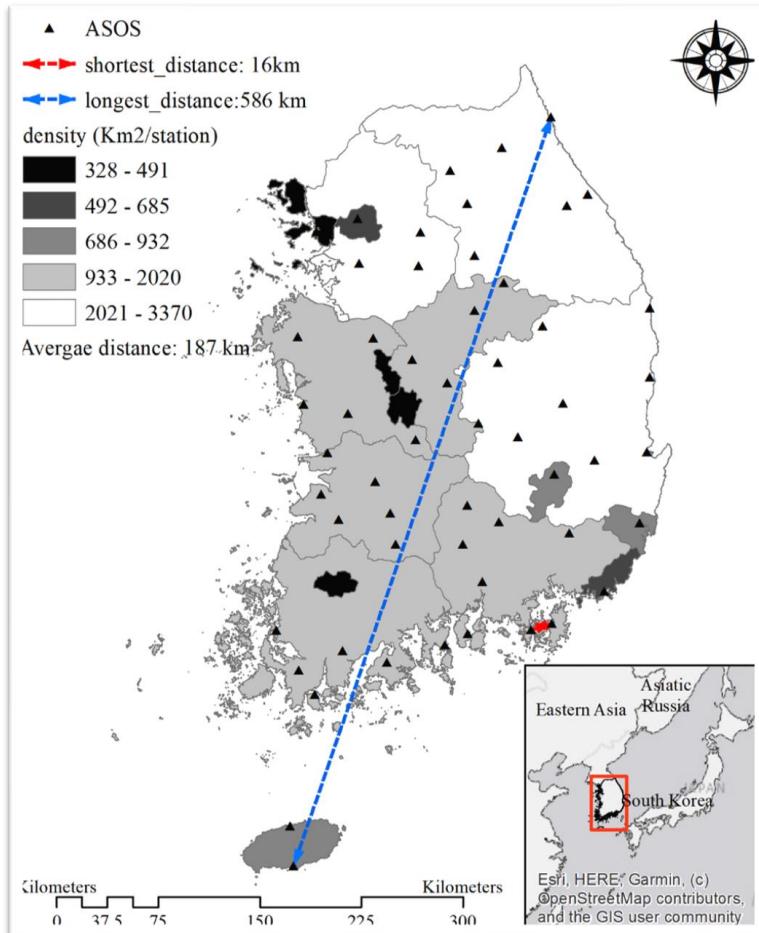


Fig 133. ASOS locations and study area

The Fig 2 shows the longest Euler distances between Sokcho and Seogwipo is 586 km (blue line). The shortest Euler distance between Geoje and Tongyeong is 16 km (red line). Average distance between two stations is 197km. The density of station of all cities and provinces is defined as the

area ( $\text{km}^2$ /one station). The minimum average  $328 \text{ km}^2$  has a station. In the East and North-East need more than  $2021 \text{ km}^2$  for a station. Therefore, the kriging model is used to predict the NDI at location without computation.

#### *|Temporal data explore*

We plot time series of Seosan (Fig 3), Inje (Fig 4) where have the long last drought duration and the most severe drought for checking. From Fig 3 and Fig 4 show various of among NDI with temporal scale 1-, 3-, 6- months. The average cumulative time for specific period leads the values are bias. The higher temporal scale like to be higher in range of NDI (NDI 6). While NDI 1 has the range of NDI is lowest.

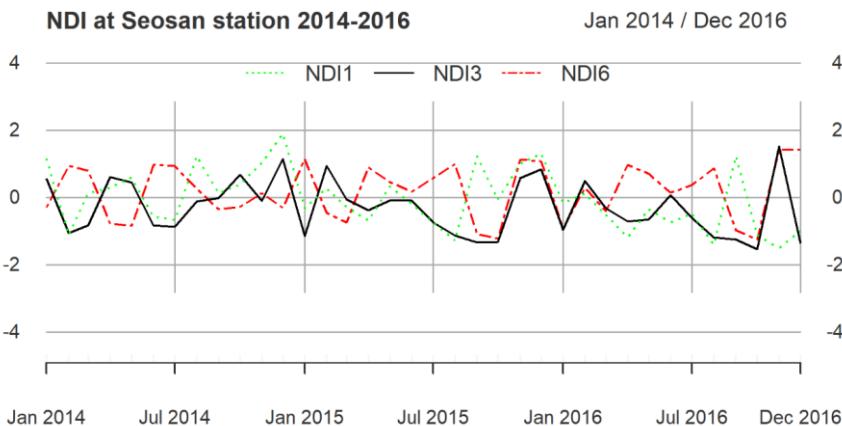


Fig 134. NDI with various temporal scale at Seosan station.

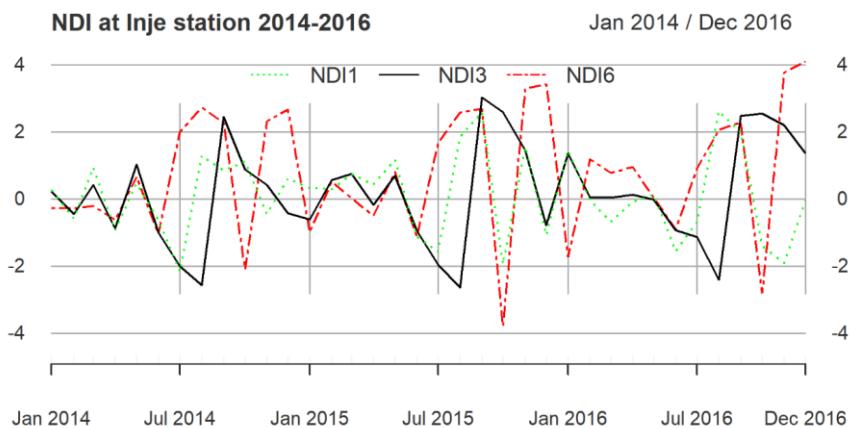


Fig 135. NDI with various temporal scale at Inje station.

Hydrology is usually changing with a cycle period. The monthly, seasonal, and yearly frequency were chosen for detecting the change. In Fig 5 presents the statistical overview NDI of each month for 36 years from 1981 to 2016 at Inje station.

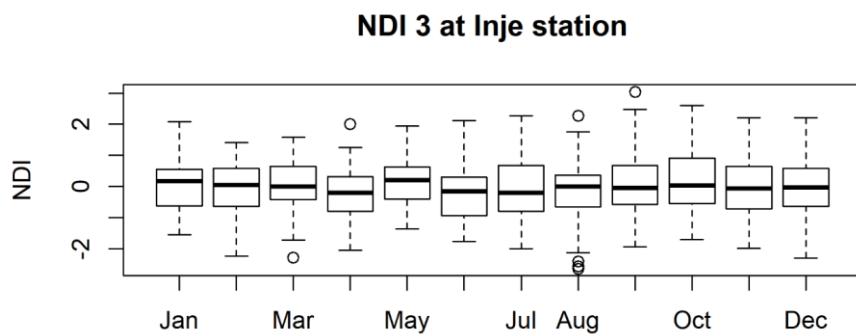


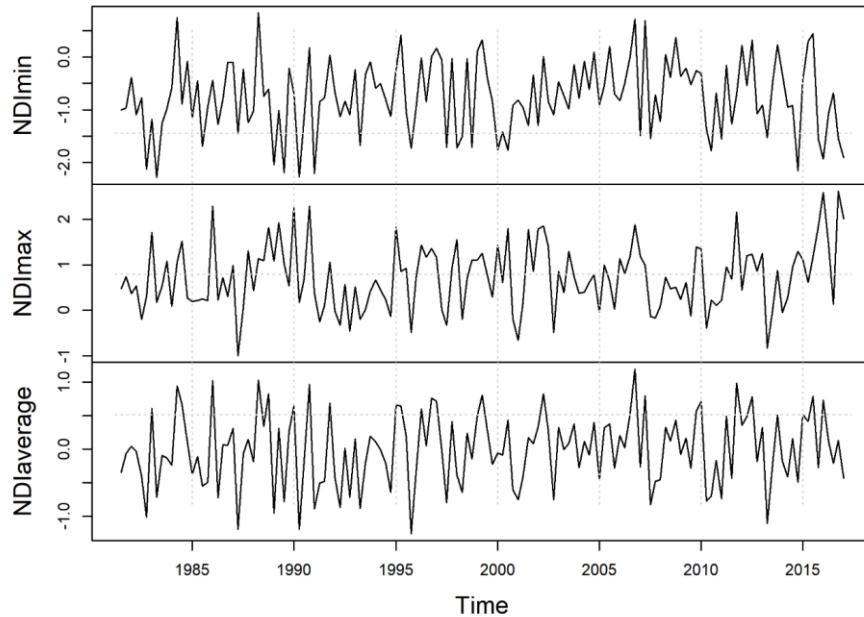
Fig 136. NDI 3 at Inje station from 1981 to 2016.

The Fig 5 shows that droughts have occurred almost every months of the years with the minimal NDI below -1. In May drought is usually less than other months. While December and August

have minimal value of NDI below -2 that imply to extreme drought. In August several extreme events had took place.

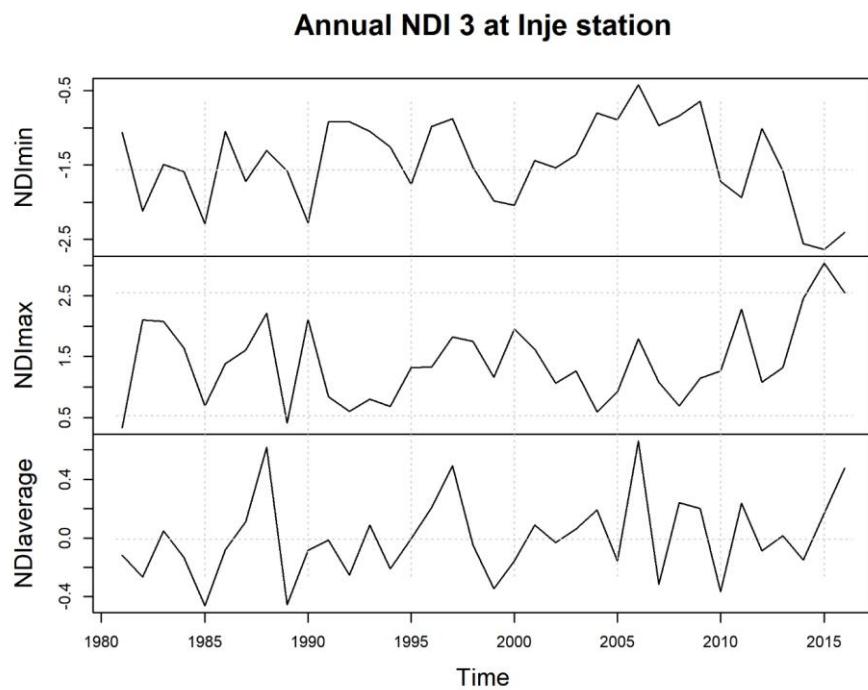
South Korea has four distinct seasons: spring (March-May), summer (June-August), fall (September-November), and winter (December-February). In Fig 6 presents the maximum, minimize, and average seasonal NDI at Inje station.

### Seasonal NDI 3 at Inje station



*Fig 137. Seasonal NDI 3 at Inje station.*

In the seasonal scale, precipitation, one important characteristic of drought was figured out that have not existence correlation to ENSO (Ho et al., 2016). In this study we would like to examine if it is having the correlation of drought and ENSO in both seasonal and yearly temporal scale. It is helpful to explain the trend of drought under the global climate. Fig 7 shows the fluctuation of maximum, minimize, and average annual NDI at the Inje stations.



*Fig 138. Minimum, maximum, and average annual NDI 3 at Inje station.*

The annual temporal shows the change of NDI 3 is clearer than seasonal, and monthly temporal scale. For instance, from 2006 to 2015 the minimal annual NDI is sharply decease. It shows the extreme drought has the increase trend in this period.

## 2. Discussion

- (1) The first steps of our framework were completed. The NDI at 59 ASOS were computed in 1-, 3-, 6- months temporal scale. Maximum, minimum, and average in monthly, seasonal, and annual temporal scale were derived at all stations. Some basic spatial distance, density of drought was computed. In the next step, we propose to create the surface model by Geostatistical approach. The most challenge that is determining

model's parameters. In case, the time is limited, we propose using a deterministic method such as inverse distance weight (IDW) for spatial interpolation.

(2) The temporal trend drought will be analyzed by NDI at 1-, 3-, 6- temporal scales. It is going to be examined in the mean, maximum, minimum at monthly, seasonal, and annual temporal scale. We chose the Inje, Seosan station as the typical demonstrations. It is hard to see if all stations were drawn in an article.

### **References**

Ho, C.-H., Choi, W., Kim, J., Kim, M.-K., Yoo, H.-D., 2016. Does El Niño-Southern Oscillation affect the precipitation in Korea on seasonal time scales? *Asia-Pacific J. Atmos. Sci.* 52, 395–403.

**WEEK 43**  
**Recent extreme drought trend in South Korea**

**3. Kriging model**

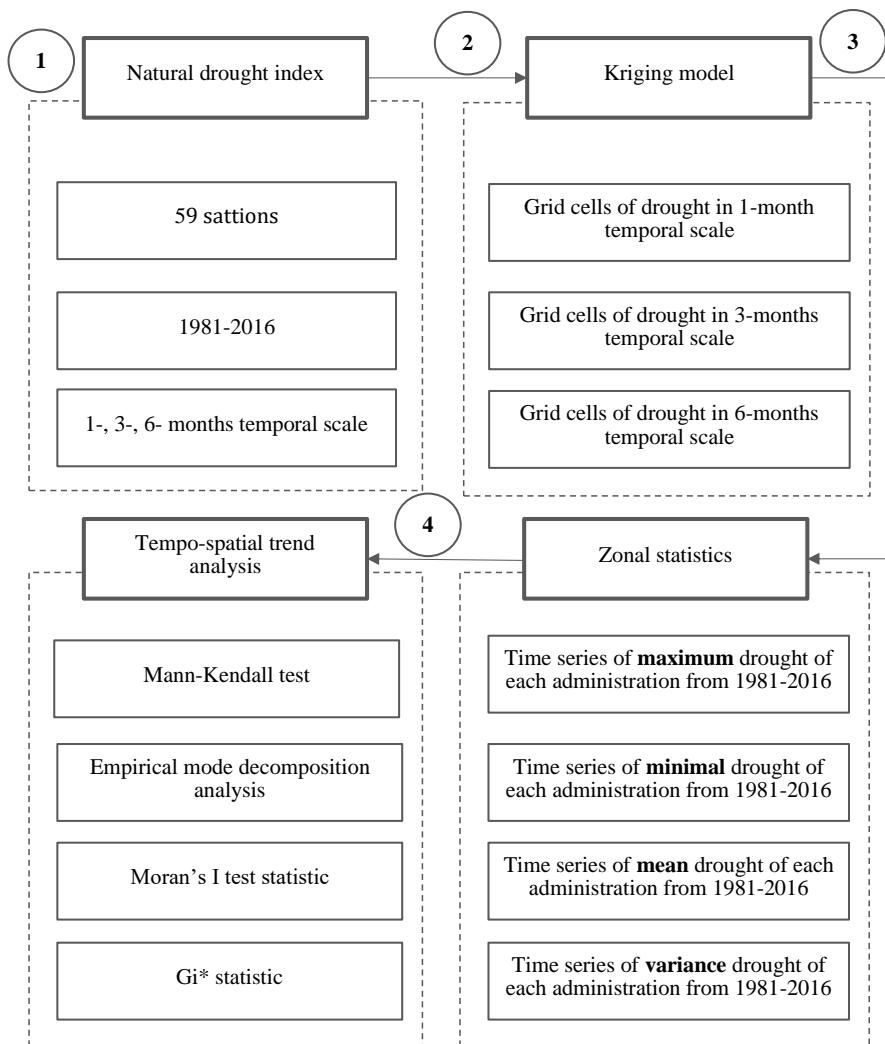


Fig 139. Flowchart for extreme drought trend assessment.

Ordinary Kriging (OK) was proposed to build the continuous drought surface for South Korea. However, determining parameters of OK model is time-consuming. Therefore, we used deterministic method as inverted distance weight (IDW). 432 maps of NDI from 1981 to 2016 were computed. The new spatial resolution was set as approximated 0.01 degree  $\sim$  10 km. The temporal scale was chosen at monthly scale. They were fitted to envelop that covers whole South Korea. One extreme drought event at September 2015 is presented in Fig 2.

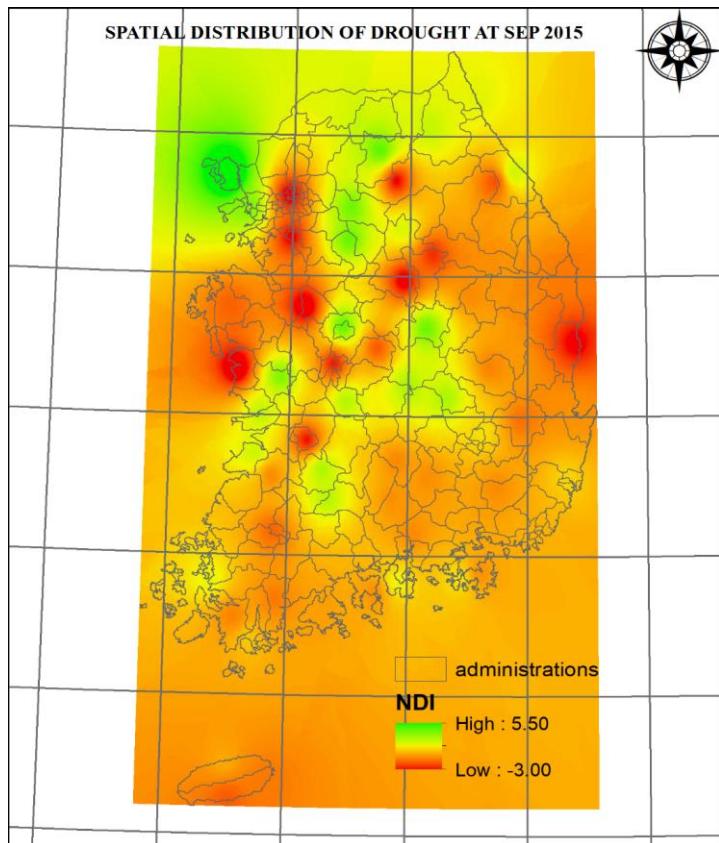


Fig 140. Spatial drought using NDI in September 2015.

#### 4. Zonal statistic

268 zones were constructed based on the administrator boundary (cities, provinces). Some redundant region has area below 10 square kilometers were neglected. In Fig 3, Fig 4 presents location and time series comparison of drought from the North of South Korea and the South of Korea.

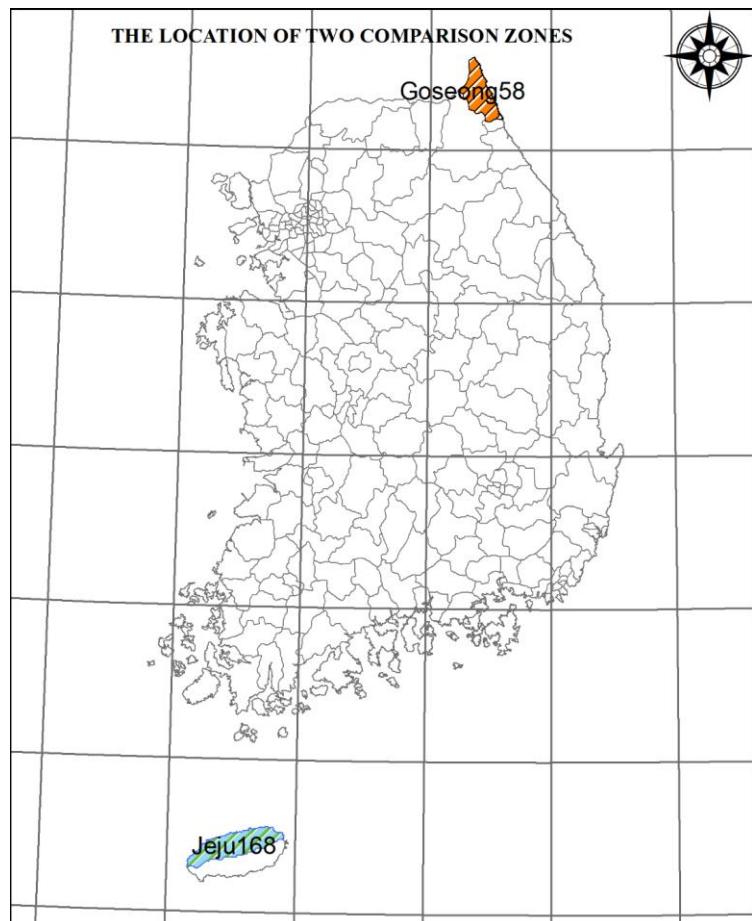


Fig 141. Location of comparison zones.

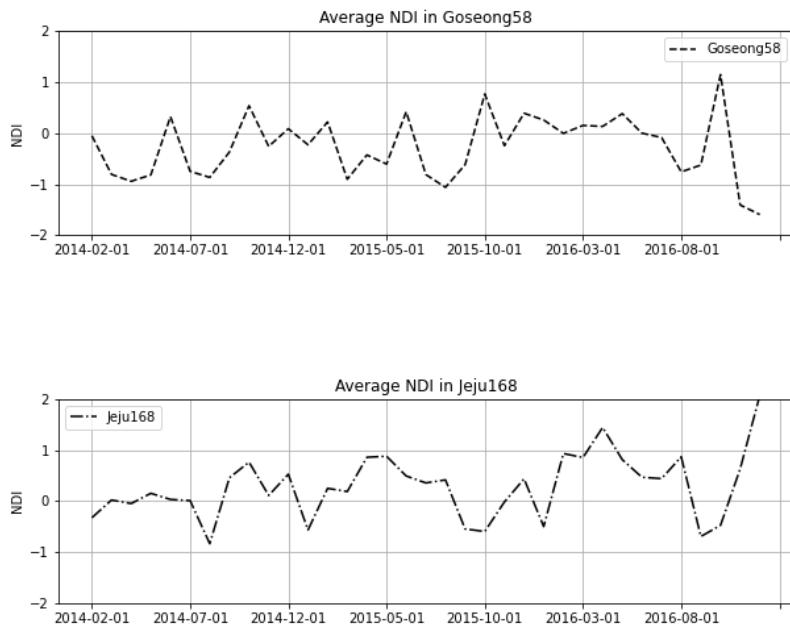


Fig 142. Average NDI from the North (Geoseong58) to the South (Jeju168) of South Korea.

##### 5. Mann Kendal

We propose using 11 types of Mann Kendal (MK) and modified methods to test the trend of drought in South Korea. Fig 5 present Hamed and Rao Modified MK test.

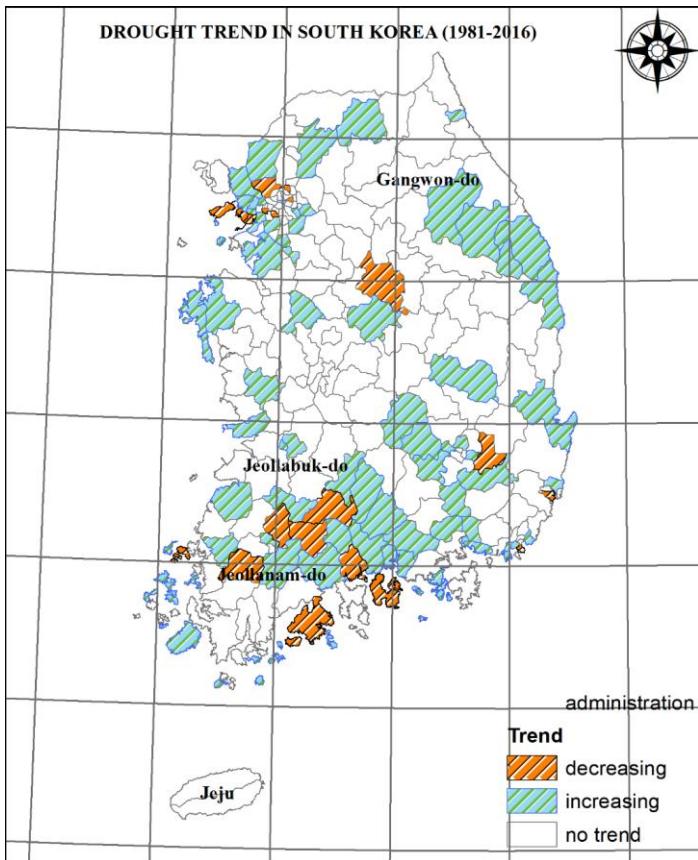


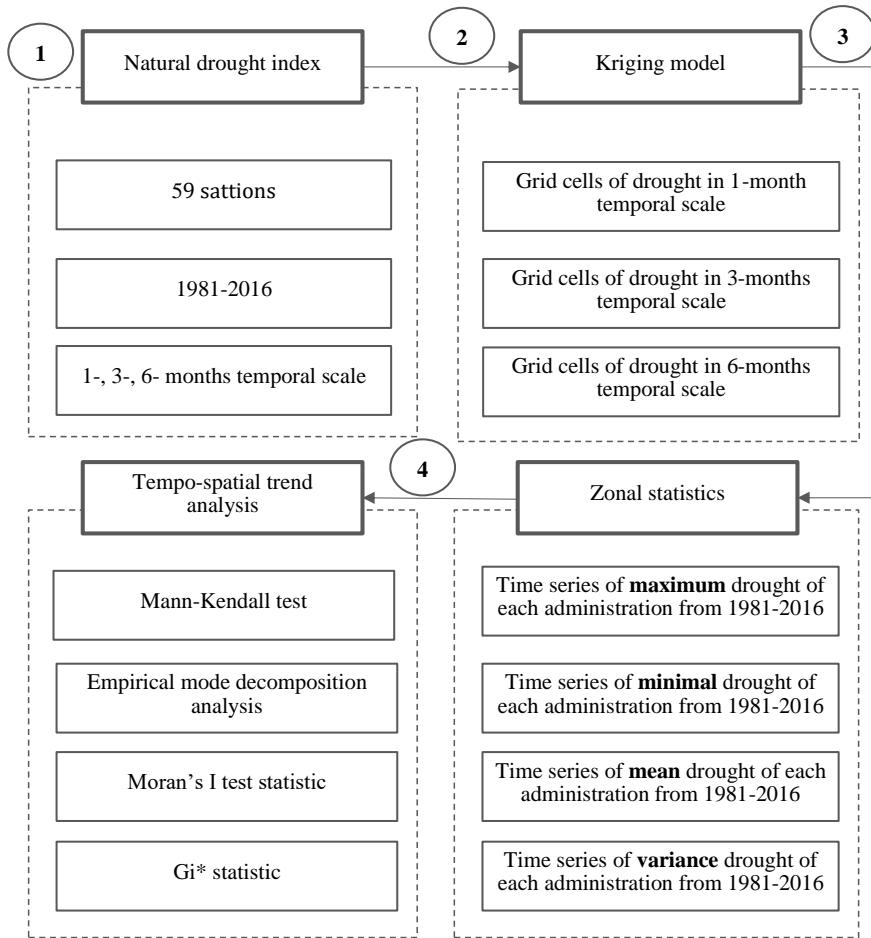
Fig 143. Trend of drought in South Korea.

## 6. Discussion

The primary results of our study were obtained. In the next steps we extended the framework with:

- (1) Analyze framework with NDI 1-months, 6-months temporal scale.
- (2) Analyze framework with maximum, minimum zonal statistic.
- (3) Analyze follow 5 years periods.
- (4) Analyze trend with other kind of MK tests.

**WEEK 44**  
Recent extreme drought trend in South Korea



#### 7. Framework estimate the trend of extreme drought

The aim of this study is estimating the drought trend. We extended computation with percentile of zones. 10 percentile values were computed as extreme values.

*Fig 144. Flowchart for extreme drought trend assessment.*

## 8. Methodology

### 2.1. Mann Kendall trend

Homogeneous test

$$Z_R = \frac{R - \frac{2N_1 N_2}{N_1 + N_2} + 1}{\sqrt{\frac{2N_1 N_2 (2N_1 N_2 - N)}{N^2 (N-1)}}} \quad (35)$$

Where  $Z_R$  is run homogeneous test result,  $R$  is run number,  $N_1$  is number of values lower than medium,  $N_2$  is number of values higher than median. If  $Z_R$  value corresponds to 5% significance level or below then the data is non-homogenous. The only homogeneous data are used to determine trend conditions.

Mann Kendall test was computed follow equation 2 to equation 5 below.

$$Z_{MK} = \begin{cases} \frac{S-1}{V(S)} & \text{for } S > 0 \\ 0 & \text{for } S = 0 \\ \frac{S+1}{V(S)} & \text{for } S < 0 \end{cases} \quad (36)$$

$$V(S) = \frac{n(n-1)(2n+5)}{18} \quad (37)$$

$$S = \sum_{k=1}^{n-1} \sum_{j=k+1}^n \operatorname{sgn}(x_j - x_k) \quad (38)$$

$$\operatorname{sgn}(x_j - x_k) = \begin{cases} +1 & \text{if } (x_j - x_k) > 0 \\ 0 & \text{if } (x_j - x_k) = 0 \\ -1 & \text{if } (x_j - x_k) < 0 \end{cases} \quad (39)$$

The  $Z_{MK}$  has positive value means increasing trend, and negative value means the decreasing trend, zero value means no trend.  $V(S)$  is the variance and  $S$  is the Kendall sum statistic. The difference between each consecutive value are computed as positive (+1), negative (-1) and neutral (0).  $x_j$  and  $x_k$  are value of time series at time  $j$ , at time  $k$  of  $n$  observational values. The trend is

considered significant if  $Z_{MK}$  is more than the significant levels. For instance, confidence 90% has  $\alpha = 10\%$  ( $Z_{MK} \geq Z_{\alpha/2} = \pm 1.645$ ).

Temporal scale 1-, 3-, 6-months NDI

Mean, max, min, standard deviation, extreme

-Monthly

-Seasonally

-Annually

-Five years period return

## 2.2. Empirical model decomposition

The mode decomposed from the Empirical Model Decomposition (EMD) is named as Intrinsic Mode Function (IMF). An IMF satisfies 2 conditions: (1) the difference between of extrema and the number of zero-crossings in the whole data span is less than or equal to one, and (2) the mean value of the envelope defined by the local maxima and envelope defined by the local minima is zero at any point. The processing of EMD follow these steps:

- a. Initialize:  $i=1$  and define  $r_0 = x(t)$ .
- b. Identify all local extrema (both maxima and minima) of  $r_0$ , connect all local maxima (minima) with curve-fitting method s the upper (lower) envelope. The cubic spline method is used to connect of extrema.
- c. Calculate the local mean curve  $m_{(j=1)}$  of the upper and lower envelops obtained from step b.
- d. Obtain the first component  $h_{(j=1)} = r_0 - m_{(j=1)}$ .
- e. Treat  $h_{(j=1)}$  as  $r_0$  and repeat steps (b) to (c) with  $j = j + 1$  as multiple times until the envelopes being symmetric to zero. The final  $h_{(j)}$  is designated as  $C_i$ .
- f. Define  $r_0 = x(t) - C_i$  and  $i = i + 1$ , and repeat steps from (b) to (e). When  $i = N$  and the residual become a monotonic function which no more IMF can be extracted, a complete shift process of series  $x(t)$  by EMD stops. The results of the EMD can be presented:

$$x(t) = \sum_{i=1}^N C_i + R_N \quad (40)$$

Where N is the number of IMFs separated from series  $x(t)$  and  $C_i$  is the ith of IMF.

### | 2.3. Moran I statistic

The degree of spatial dependence in trend of NDI was determined among 269 subdistricts based on Moran's I test statistic. The global measure of spatial autocorrelation is calculated as follows:

$$I = \frac{n}{\sum_{i=1}^n (\bar{NDI}_i - \bar{NDI})^2} \frac{\sum_{i=1}^n \sum_{j=1}^n W_{ij} (\bar{NDI}_i - \bar{NDI})(\bar{NDI}_j - \bar{NDI})}{\sum_{i=1}^n \sum_{j=1}^n W_{ij}} \quad (41)$$

$\bar{NDI}_i$  and  $\bar{NDI}_j$  refer to the natural drought index trend in station i and station j, respectively; n is the number of spatial units indexed by i and j; NDI is the overall mean

of the natural drought index.  $W_{ij}$  is a matrix of spatial weights, that is if station i and station j are adjacent  $W_{ij} = 1$  otherwise  $W_{ij} = 0$ .

### 2.4. Gi\* statistic

$Gi^*$  statistic was used to estimate the spatial extreme drought patterns (Ord and Getis 1995). The Z score output of  $Gi^*$  showed features with high or low values that had a clustered trend pattern.

Local statistics of hotspot analysis are given as follows:

$$Z(G_i^*) = \frac{\sum_{j=1}^n w_{i,j} x_j - \bar{X} \sum_{j=1}^n w_{ij}^2}{S \sqrt{\left[ n \sum_{j=1}^n w_{i,j}^2 x_j - (\sum_{j=1}^n w_{i,j} x_j)^2 \right] / (n-1)}} \sim N(0,1) \quad (42)$$

where  $x_j$  is the attribute value for feature  $j$ ,  $w_{i,j}x_j$  is the spatial weight between feature  $i$  and  $j$ ,  $n$  is equal to the total number of features.

$$\bar{X} = \frac{\sum_{j=1}^n x_j}{n} \quad (43)$$

$$S = \sqrt{\frac{\sum_{j=1}^n x_j^2}{n} - (\bar{X})^2} \quad (44)$$

The  $Gi^*$  value in Equation (8) is normalized to the  $Z$  value. If the  $Z$  score of  $Gi^*$  is greater than 1.65, the region can be considered a hot spot at a confidence level of 90%. If the  $Z$  score of  $Gi^*$  is lower than 1.65, the region can be considered a cold spot at a confidence level of 90%. If the  $Z$  score of  $Gi^*$  is larger than 1.96, the region can be considered a hot spot at the confidence level of 95%. If the score is less than -1.96, the region is a cold spot at the same confidence level.

## 9. Results

### 3.16. Drought trend using Mann Kendall test

The trend of drought was examined with 1-, 3-, 6- months scales. Several characteristics of zones such as min value, max value, standard deviation, average value, and extreme value at 10 percentiles were analyzed to figure out adequate drought properties. Fig 2 to Fig 6 present results of the original Mann Kendall test.

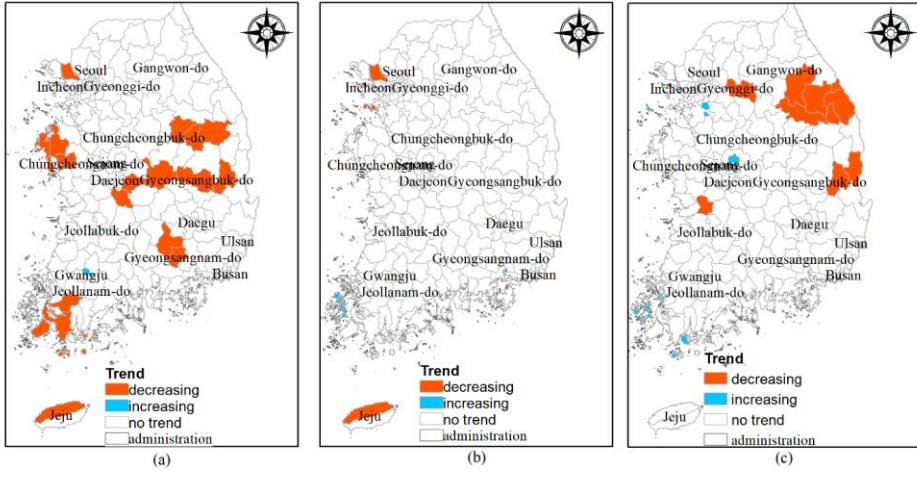


Fig 145. Trend of zonal statistic minimal NDI 1-month (a), NDI 3-months (b), and NDI 6- months temporal scales.

For convenient NDI 1-, 3-, 6- months temporal scale was named as NDI 1, NDI 3, and NDI 6. Based on Fig 2, NDI 3 find the most trend of zones compares with NDI 3 and NDI 6. However, NDI 6 found the biggest subdistrict has NDI trend decreasing.

The maximum of NDI corresponds to the less drought sever situation. Fig 3 shows locations that maximum NDI has increasing trend. In this case, NDI 3 explores several locations has NDI increasing trend.

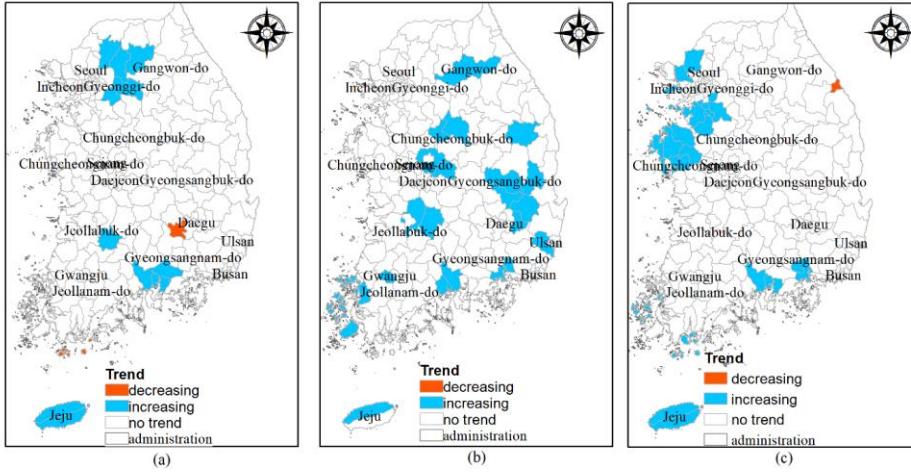


Fig 146. Trend of zonal statistic maximal NDI 1-month (a), NDI 3-months (b), and NDI 6-months temporal scales.

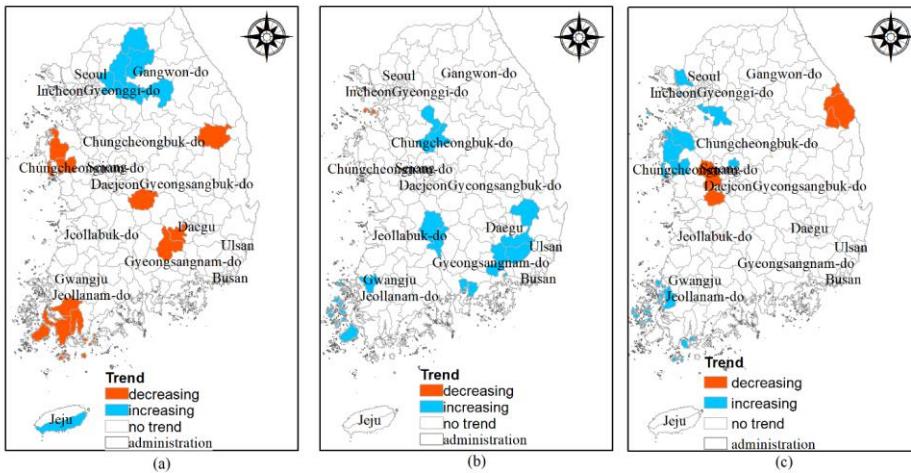


Fig 147. Trend of zonal statistic average NDI 1-month (a), NDI 3-months (b), and NDI 6-months temporal scales.

Fig 4 shows that NDI 1 reveals the most locations has decreasing, or increasing trend compare to NDI 3, NDI6.

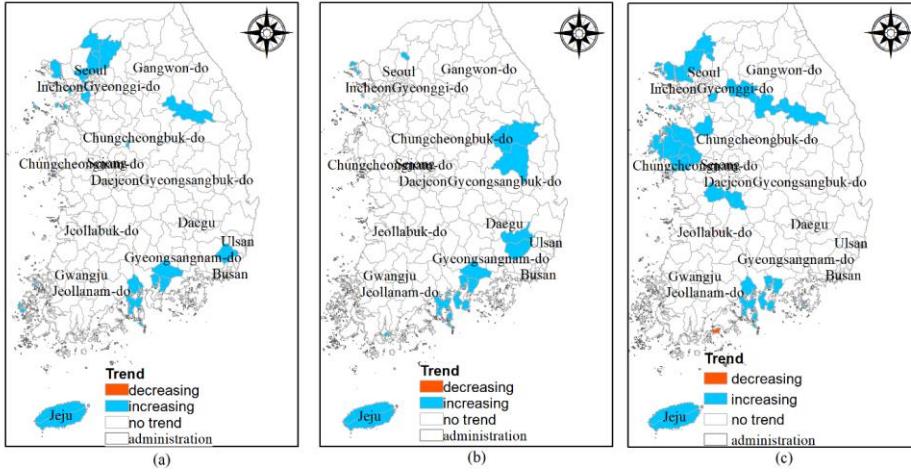


Fig 148. Trend of zonal statistic standard deviation NDI 1-month (a), NDI 3-months (b), and NDI 6-months temporal scales.

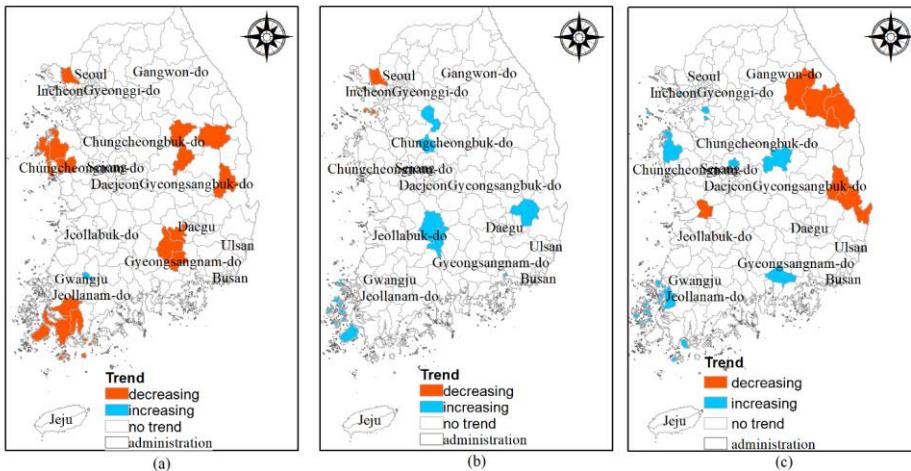


Fig 149. Trend of zonal statistic extreme (10 percentile) NDI 1-month (a), NDI 3-months (b), and NDI 6-months temporal scales.

NDI 1 determined the most locations have decreasing trend.

3.17. *Drought trend using Empirical Model decomposition*

3.18. *Drought trend using Moran I test*

3.19. *Drought trend using Gi\* statistic*

## 10. Discussion

(3) The result should be selected. We could not give all of results in our frameworks. Reference to other articles to choose the suitable results. We proposed presentation the trend of drought using Mann Kendall test with multiple temporal scales and zonal statistic. We got the most important data are time series of each zone. From this data we can derive to seasonal, annual, multiple years period return to estimate trends. The other modified Mann Kendall test, the Empirical Model Decomposition, and seasonal, multiple years period, correlation to climate index such as NINO will be research after conferences. In addition, using this data, the study can be future developed. Using Deep Learning to predict extreme drought change could be interesting for our next studies.

(4) The computational time of derive zonal time series for is lengthy. We need extract time series for 269 zones from images. Each image has  $250 \times 380$ , total 895,00 pixels. We must compute 432-time steps corresponding to the number months from 1981 to 2016. The zonal statistic consists of 5 types: max, min, average, standard deviation and extreme. All of processing above were analyzed follow 4 types of temporal scales: 1 months, 3 months, 6 months. Therefore, we need to handle  $269 \times 89,500 \times 432 \times 5 \times 3 = 1.56 \times 10^{12}$  calculations (approximates 1,560 billion computational steps). It should be improved by parallel computation in the future.

Tuong comment:

-Computed trend of extreme 6 months with moonsoon and years data

## WEEK 45

### Recent extreme drought trend in South Korea

#### 11. Framework estimate the trend of extreme drought

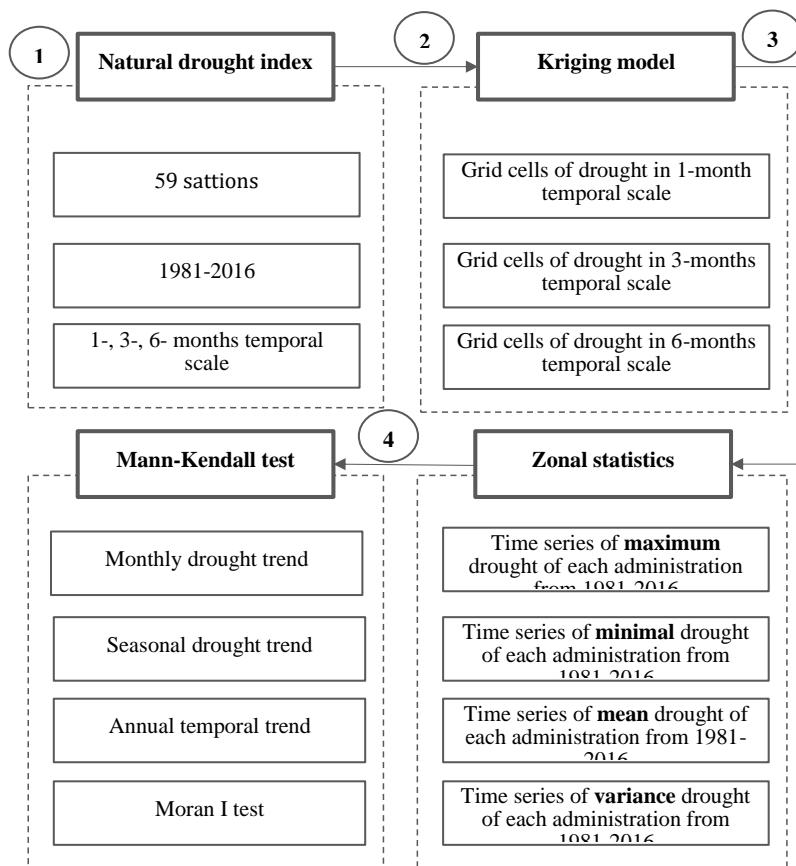


Fig 150. Flowchart for extreme drought trend assessment.

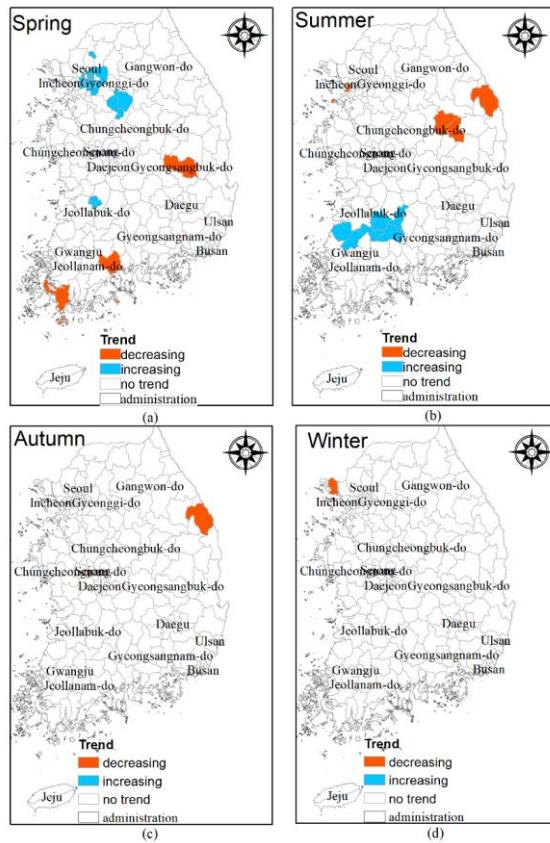
The aim of this study is estimating the drought trend under multiple time and spatial scale. The Geostatistics, Zonal statistic and non-parametric trend test were utilized to detected spatial and temporal extreme drought trend.

## 12. Extended analysis

### 2.5. Seasonal drought trend

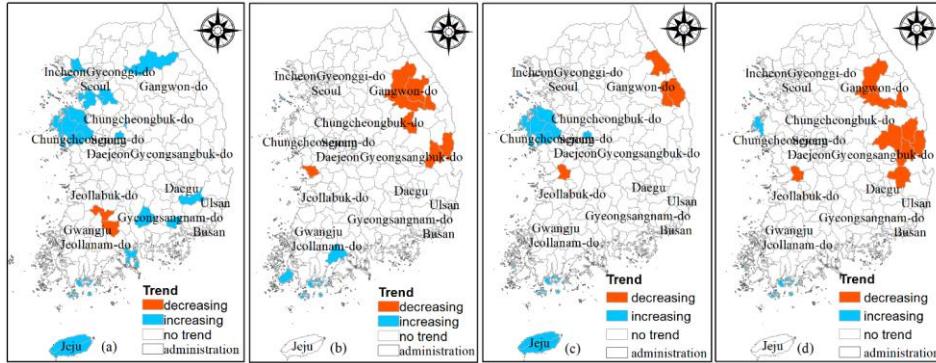
The extreme drought in Summer has increasing trend. Although, it is a wet season, the high temperature lead to increase evapotranspiration. It is the reason that in Summer, extreme drought trend is increasing (Fig 2).

### 2.6. Annual drought trend



*Fig 151. Extreme drought trend follows seasons.*

At the annual temporal scale, drought has increasing trend at maximum, minimum, average, and extreme. The annual extreme drought has significantly increasing compare with others (Fig 3). In the North East of South Korea is the location where drought trend is level up.



*Fig 152. Annual drought trend with maximum (a), minimum (b), average (c) and extreme (d) statistics.*

## 2.7. Ten types of Mann Kendall test

10 types of Man Kendall family test were examined in this study.

- Original Mann-Kendall test (Kendall, 1975; Mann, 1945) is a nonparametric test, which does not consider serial correlation or seasonal effects.
- Hamed and Rao Modified MK Test is the method that modified MK test by Hamed and Ramachandra Rao (1998) to address serial autocorrelation issues. They suggested a variance correction approach to improve trend analysis.
- Yue and Wang Modified MK Test is method that is also variance correction method for considered serial autocorrelation proposed by (Yue & Wang, 2004).
- Modified MK test using Pre-Whitening method was suggested by (Yue et al., 2002) to use Pre-Whitening the time series before the application of trend test.
- Modified MK test using Trend free Pre-Whitening method was proposed by (Yue et al., 2002) to remove trend components and then Pre-Whitening the time series before application of trend test.

- f) Multivariate MK Test was proposed by Hirsch et al. (1982). They used this method for seasonal MK tests, where they considered every month as a parameter.
- g) For seasonal time series data, (Hirsch et al., 1982) proposed Seasonal MK Test to calculate the seasonal trend.
- h) Based on the proposed seasonal MK test of Hirsch et al. (1982), (Helsel & Frans, 2006) suggest a regional MK test to calculate the overall trend on a regional scale.
- i) Correlated Multivariate MK Test was proposed by Hipel and McLeod (1994), for when time series significantly correlate with the preceding one or more months/seasons.
- j) In a real event, many factors affect the main studied response parameter, which can bias the trend results. To overcome this problem, Libiseller and Grimvall (2002) proposed this partial MK test. It required two parameters as input, where one is the response parameter and other is an independent parameter.

Table 21. Result from first 7 tests for 15 sub-districts

sta	origi nal_t est	hamed_rao_ modification _test	yue_wang_ modificatio n_test	pre_whitenin g_modificatio n_test	trend_free_pre_w hitening_modificat ion_test	multiv ariate_ test
Ando ng10 6	no trend	no trend	increasing	no trend	no trend	no trend
Ansa n75	no trend	no trend	increasing	no trend	no trend	no trend
Anso eng7 6	no trend	no trend	increasing	no trend	no trend	no trend
Anya ng77	no trend	no trend	increasing	no trend	no trend	no trend
Asan 25	no trend	no trend	increasing	no trend	no trend	no trend

sta	origi nal_t est	hamed_rao_ modification _test	yue_wang_ modification _test	pre_whitenin g_modificatio n_test	trend_free_pre_w hitening_modificat ion_test	multiv ariate_ test
Boeu n13	no trend	no trend	increasing	no trend	no trend	no trend
Bong hwa1 07	no trend	no trend	increasing	no trend	no trend	no trend
Bory eong 26	no trend	no trend	increasing	no trend	no trend	no trend
Bose ong1 86	no trend	no trend	increasing	no trend	no trend	no trend
Buan 170	no trend	no trend	increasing	no trend	no trend	no trend
Buan 171	no trend	no trend	increasing	no trend	no trend	no trend
Buch eon7 8	no trend	no trend	increasing	no trend	no trend	no trend
Buk0	no trend	no trend	increasing	no trend	no trend	no trend
Buk2 64	no trend	no trend	no trend	no trend	no trend	no trend
Buk4 1	no trend	no trend	decreasing	no trend	no trend	no trend
Buk7 2	no trend	no trend	no trend	no trend	no trend	no trend

## 2.8. Moran I

The spatial autocorrelation shows various cluster of droughts. However, the extreme drought trend was not located at the region has the clustering pattern (Fig 4). The cluster was test with confident level at 95%.

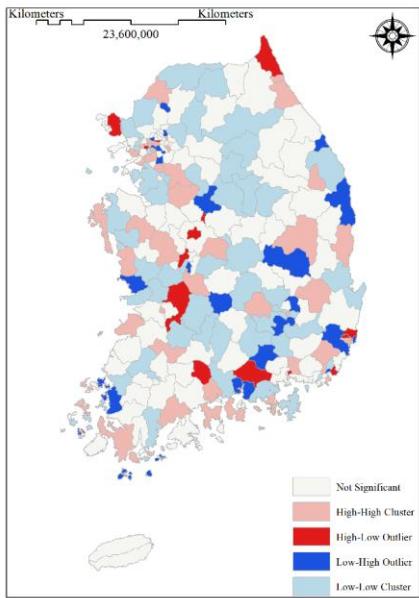


Fig 153. Spatial autocorrelation of extreme drought.

### 13. Discussion

- (1) The initial results of our framework were obtained. We propose find the uniqueness, rearrangement to be a completed study.
- (2) The study about using NDI, Satellite data, and ANN to improve spatial drought coverage is doing simultaneously.

### References

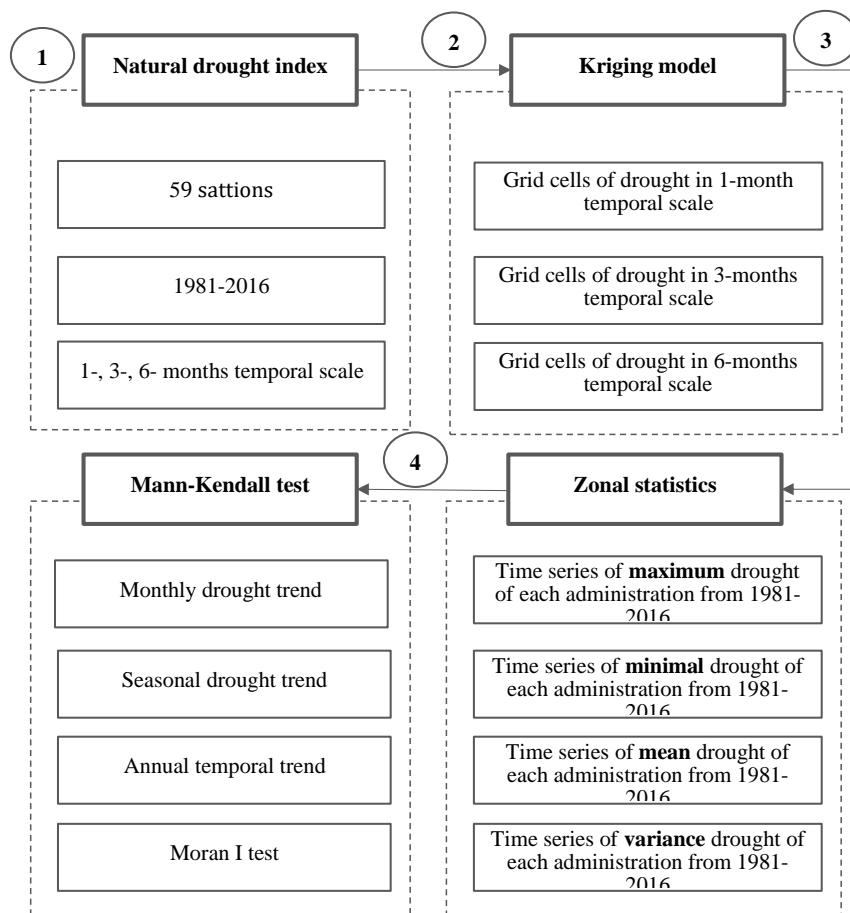
- Hamed, K. H., & Ramachandra Rao, A. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, 204(1-4), 182-196. doi:10.1016/S0022-1694(97)00125-X
- Helsel, D. R., & Frans, L. M. (2006). Regional Kendall Test for Trend. *Environmental Science & Technology*, 40(13), 4066-4073. doi:10.1021/es051650b
- Hipel, K. W., & McLeod, A. I. (1994). *Time series modelling of water resources and environmental systems*: Elsevier.
- Hirsch, R. M., Slack, J. R., & Smith, R. A. (1982). Techniques of trend analysis for monthly water quality data. *Water resources research*, 18(1), 107-121. doi:10.1029/WR018i001p00107

- Kendall, M. (1975). Rank correlation measures *London: Charles Griffin*, 220.
- Libiseller, C., & Grimvall, A. (2002). Performance of partial Mann-Kendall tests for trend detection in the presence of covariates. *Environmetrics*, 13(1), 71-84. doi:10.1002/env.507
- Mann, H. B. (1945). Nonparametric Tests Against Trend. *Econometrica*, 13(3), 245. doi:10.2307/1907187
- Yue, S., Pilon, P., Phinney, B., & Cavadias, G. (2002). The influence of autocorrelation on the ability to detect trend in hydrological series. *HYDROLOGICAL PROCESSES*, 16(9), 1807-1829. doi:10.1002/hyp.1095
- Yue, S., & Wang, C. (2004). The Mann-Kendall Test Modified by Effective Sample Size to Detect Trend in Serially Correlated Hydrological Series. *Water resources management*, 18(3), 201-218. doi:10.1023/B:WARM.0000043140.61082.60

## WEEK 46

### Recent extreme drought trend in South Korea

#### 1. Review recent drought trend



*Fig 154. Flowchart for extreme drought trend assessment.*

The aim of this study is estimating the drought trend under multiple time and spatial scale. The Geostatistics, Zonal statistic and non-parametric trend test were utilized to detected spatial and temporal extreme drought trend.

Using results from RCP 8.5 climate change scenarios, Park et al. (2015) estimate drought in South Korea until 2100. Daily Effective Drought Index (EDI) was used to predict drought. 678 grid points (12.5 km interval) were grouped into 4 groups based on the similar climate. G1 (Northwest), G2 (Middle), G3 (Northeast), and G4 (Southern) regions were constructed. Future drought events are quantified and ranked depending on the duration and intensity. It figured out (when, where, how severe) drought within the clustering regions above. Linear regression was used in that study. Several limitation of linearity regressions such as: based on the stationary assumption, sensitive with extreme values.

The seasonal drought variation and trend over central Europe were estimated in study (Hänsel et al., 2019). Seasonal climate trends from 1951-2015 was examined by using 91 climate stations. The trend of drought is significantly affected by choosing periods. The other study detects drought change by using multiple evapotranspiration was developed by Shi et al. (2020). These authors used two different emission scenarios (RCP 4.5 and RCP 8.5) to estimate drought change.

The previous studies assess the future drought based on the historical data, or simulation data from climate model. The climate model has some limitation such as it based on a assumption that climate consistency. The change of climate is constant. It is difficult to be obtained in real. The statistical analysis based on history data is normally lack of data, and it does not include the climate characteristic.

In this study, we proposed using both historical data and climate simulation data to estimate extreme drought. The assimilated data of historical and climate model are stimulant used to predict drought trend.

## 2. The spatial drought improvement

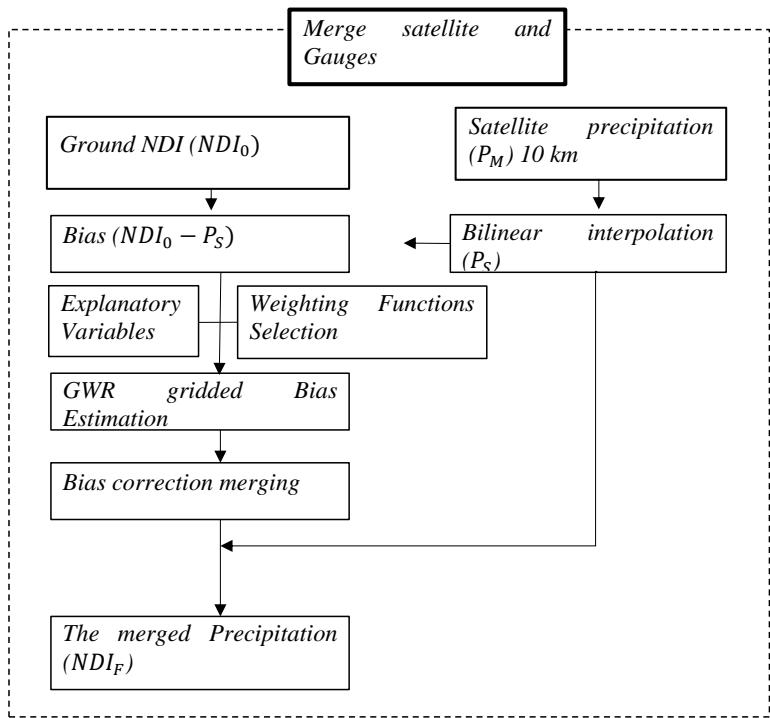


Fig 155. The flowchart of merging satellite and gauges data.

## 3. Discussion

(4) We propose make our results in detail such as classification drought in drought events. From that the drought trend can be considered drought parameters: drought duration, drought severity and drought interval. The previous studies estimate drought trend based on historical data or based on climate model. Both approaches have the positive and negative in the assumptions and results. We propose using both historical data and climate model data to estimate the long-term extreme drought. The other approach, we propose using a Deep learning approach to predict extreme

drought in future. Then, it will be compared with the statistical approach. Therefore, next step of the first study, we will select and exploit a suitable climate model. Precipitation, Runoff, Soil moisture from climate model will be used to compute the future NDI. Or using a deep learning algorithm to predict extreme drought. Finally, extreme drought trend for historical and future drought will be obtained.

(5) Review the improve spatial analysis using satellite and gauges data. For second study, next steps are setting up an ANN algorithm as the function to replace GWR. The result from ANN model will compare with original GWR to adjust the improvement of proposal framework.

(6) Compare two studies above, we thought study about estimate extreme drought trends could be completed easily. Therefore, we propose to complete it firstly.

## References

- Hänsel, S., Ustrnul, Z., Łupikasza, E., & Skalak, P. (2019). Assessing seasonal drought variations and trends over Central Europe. *Advances in Water Resources*, 127, 53-75.  
doi:<https://doi.org/10.1016/j.advwatres.2019.03.005>
- Park, C.-K., Byun, H.-R., Deo, R., & Lee, B.-R. (2015). Drought prediction till 2100 under RCP 8.5 climate change scenarios for Korea. *Journal of Hydrology*, 526, 221-230.  
doi:<https://doi.org/10.1016/j.jhydrol.2014.10.043>
- Shi, L., Feng, P., Wang, B., Liu, D. L., & Yu, Q. (2020). Quantifying future drought change and associated uncertainty in southeastern Australia with multiple potential evapotranspiration models. *Journal of Hydrology*, 590, 125394.  
doi:<https://doi.org/10.1016/j.jhydrol.2020.125394>

## WEEK 47

### Recent extreme drought trend in South Korea

#### 1. Extreme drought trend studies using statical and machine learning approach.

##### *|Drought trend using statistical method.*

Several studies used Mann-Kendall (MK) test to detect drought trends (Guo et al., 2019; Nouri & Homae, 2020; Sharma & Goyal, 2020). Decomposition ensemble model (DE) was utilized to analysis drought trend (Yang et al., 2019). Wavelet transform (WT) was chosen to assess drought trend for India (Sharma & Goyal, 2020). The Hibert-Huang transform (HHT) and DE was selected to examine the drought change. The other drought trend can be figured out by the linear regression (Chanda & Maity, 2017).

From previous studies, it found that MK is the most popular methods to determine drought trends. It can apply with various drought indices, spatial and temporal scales. Therefore, we proposed using MK as main trend test methods Drought trend using machine learning approaches.

However, MK based on stationary assumption. It is difficulty obtained. Predict the future trends based on MK could be less accurate. We proposed using Machine Learning incorporated with MK to solve this issue.

##### *|Drought trend using Machine learning methods.*

ANN, SVN were used to forecast drought trend in New South Wales, Australia (Dikshit et al., 2020). Other drought trend study (Eroğluer & Apaydin, 2020) was used ANN as the main approach to predict drought. Random forest (RF), Multiple layers perceptron (MLP), and neuro-fuzzy interface system (ANFIS) were also the widely applied method for prediction drought (Mohamadi et al., 2020; Zeraatpisheh et al., 2019).

#### 2. The uniqueness of our study

The statistical method and machine learning method were succeeded to predict drought trend. In this study, we propose a framework combine statical and machine learning method to predict drought trend. It is demonstrated in the South Korea case. We used Natural Drought Index, one integrated drought index consists of meteorological, hydrological, and agricultural drought.

##### *|Application machine learning to predict drought timeseries.*

Multiple time series are predicted by RNN to extend drought data. Fig1 and Fig 2 presents the processing of training, testing, and result of the model.

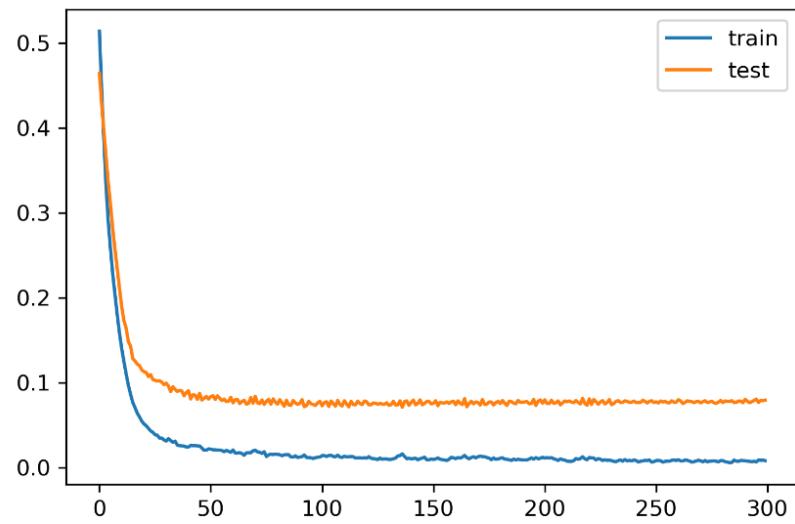


Fig 156. Training and testing for zone 1.

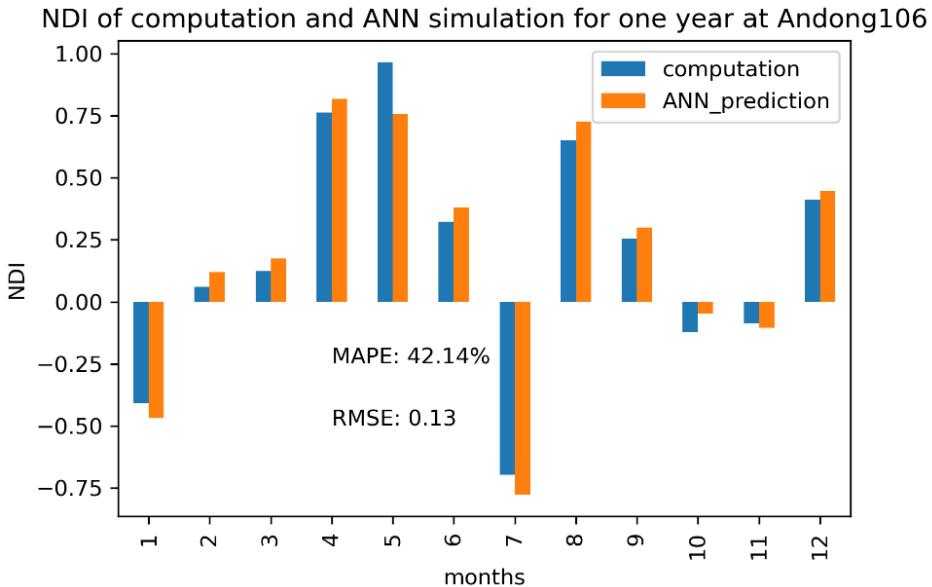


Fig 157. Compare computed NDI and ANN predicted NDI.

Based on Fig 1 we can see that our model is underfitting. The training can “learning “more until the slopes equal zero. It leads the result in Fig 2 is not high. However, the trend of ANN predicted NDI like computed NDI. Therefore, ANN predicted NDI could be used to predict extreme drought.

### 3. Discussion

In the next step, we proposed optimize parameters to improve the accurate. Complete build machine learning model for the rest of other zones. Finally, the framework of statistic and machine learning will be used to estimate and predict extreme drought trend.

### References

- Chanda, K., & Maity, R. (2017). Assessment of Trend in Global Drought Propensity in the Twenty-First Century Using Drought Management Index. *Water Resources Management*, 31(4), 1209-1225. doi:10.1007/s11269-017-1571-3
- Dikshit, A., Pradhan, B., & Alamri, A. M. (2020). Temporal Hydrological Drought Index Forecasting for New South Wales, Australia Using Machine Learning Approaches. *Atmosphere*, 11(6), 585.

- Eroğluer, T. A., & Apaydin, H. (2020). Estimation of Drought by Streamflow Drought Index (SDI) and Artificial Neural Networks (ANNs) in Ankara-Nallıhan Region. *Turkish Journal of Agriculture-Food Science and Technology*, 8(2), 348-357.
- Guo, X. L., Yang, Y., Li, Z. S., You, L. Z., Zeng, C., Cao, J., & Hong, Y. (2019). Drought Trend Analysis Based on the Standardized Precipitation-Evapotranspiration Index Using NASA's Earth Exchange Global Daily Downscaled Projections, High Spatial Resolution Coupled Model Intercomparison Project Phase 5 Projections, and Assessment of Potential Impacts on China's Crop Yield in the 21st Century. *Water*, 11(12). doi:10.3390/w11122455
- Mohamadi, S., Sammen, S. S., Panahi, F., Ehteram, M., Kisi, O., Mosavi, A., . . . Al-Ansari, N. (2020). Zoning map for drought prediction using integrated machine learning models with a nomadic people optimization algorithm. *Natural Hazards*, 104(1), 537-579.
- Nouri, M., & Homaei, M. (2020). Drought trend, frequency and extremity across a wide range of climates over Iran. *Meteorological Applications*, 27(2). doi:10.1002/met.1899
- Sharma, A., & Goyal, M. K. (2020). Assessment of drought trend and variability in India using wavelet transform. *Hydrological Sciences Journal-Journal Des Sciences Hydrologiques*, 65(9), 1539-1554. doi:10.1080/02626667.2020.1754422
- Yang, H. C., Wang, H. X., Fu, G. B., Yan, H. M., & Zhao, P. P. (2019). Evaluation of HHT approach for estimating agricultural drought trend and frequency based on modified soil water deficit index (MSWDI). *Theoretical and Applied Climatology*, 137(3-4), 1825-1842. doi:10.1007/s00704-018-2688-x
- Zeraatpisheh, M., Ayoubi, S., Jafari, A., Tajik, S., & Finke, P. (2019). Digital mapping of soil properties using multiple machine learning in a semi-arid region, central Iran. *Geoderma*, 338, 445-452. doi:<https://doi.org/10.1016/j.geoderma.2018.09.006>

## WEEK 48

### Recent extreme drought trend in South Korea

#### 4. Optimize hyperparameters

Hyperparameters optimization (HPO) is an important part for design neural network structures and model training process. Hyperparameter refers to parameters that cannot be updated during the training. They relate to structure of the model, such as the number of hidden layers, activated functions, learning rate, batch size, and optimizer. The HPO is the final step of model design and the first step of training a neural network.

##### *|Grid search*

Grid search performs an exhaustive search on the hyperparameter set specified by users. Users must have some preliminary knowledge on these hyper-parameters because it generates all candidates. Grid search is applied for several hyper-parameters with a limited search space. Grid search is the most straightforward search algorithm that leads to the most accurate predictions (Joseph, 2018). However, grid search has high dimensionality that leads computational resources increases exponentially.

##### *|Random search*

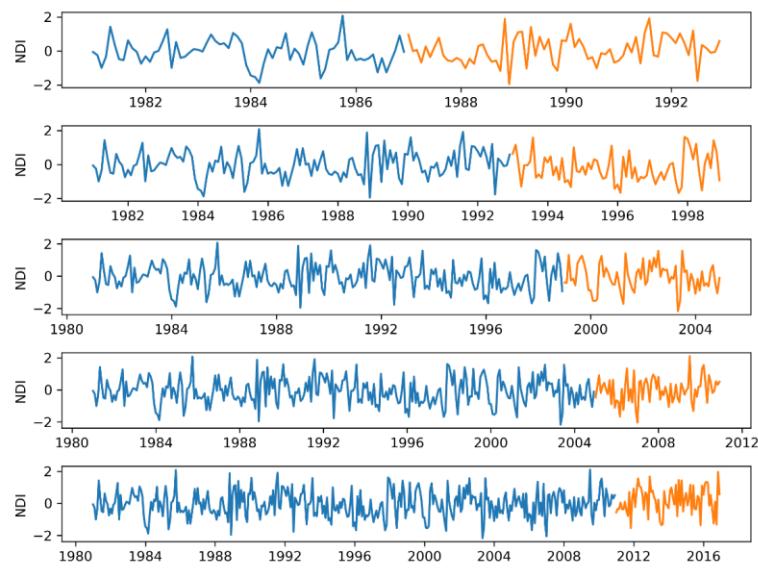
Random search (Bergstra & Bengio, 2012) is a basic improvement on grid search. It indicates a randomized search over the hyperparameters from certain distributions over possible values. The searching continues until the desired accuracy is reached. Random search has to benefit: first: the search space can be assumed independently follow to the distribution; seconds, it is less time consuming.

##### *|Bayesian optimization*

Bayesian optimization was proposed by Močkus (1975), and later became popular when it was applied to the global optimization problems (Jones et al., 1998). Bayesian optimizes hyperparameter is aim to ovoid local minimal. This method finds the global optimum with minimum number of trials. Bayesian has advantages: first, users are not required to possess preliminary of the distribution of hyper-parameters, and the second, it is using a posterior probability.

#### 5. Improve the method for model estimation

Before optimizing hyperparameters, we improve the method of evaluation. We used the cross validation (CV) for sequence data. CV for sequence data does not split random dataset as K-fold cross validation, which is used popular in machine learning. K-fold cannot be directly used because they assume that there is no relationship between the observations. Each observation is independent. That is not true for time series data as NDI. Instead, we split the data up and respect the temporal order in which values were observed. Fig 1 present CV with 5 split parts for Ansan75 zone.



*Fig 158. CV with 5 split parts for Ansan75 zone.*

Using CV can overcome result based on limited events. The model is evaluated with 5 periods with various ratios of training and testing. Fig 2 presents the absolute percentage error of the model.

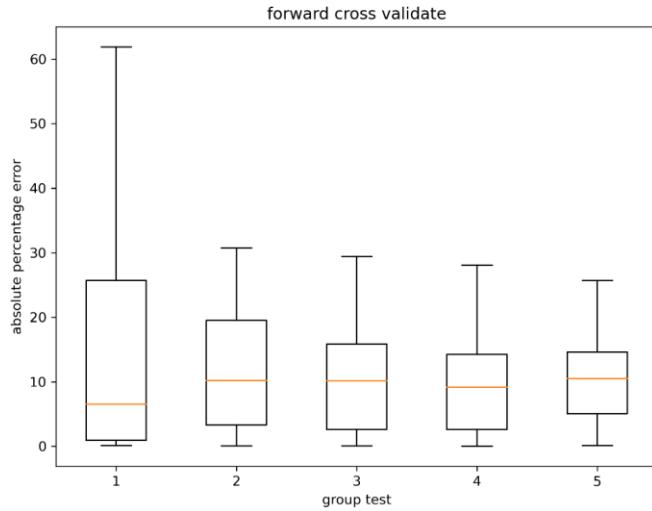


Fig 159. Absolute percentage error for 5 cross validation.

Based on Fig 2, the mean of absolute percentage error of group test 1 is lowest. However, it is uncertainty with the extreme error is over 60 %. It is reasonable because this group test corresponding to the lowest ratio train / test data. Similarly, group test 5 with the highest ratio of training / testing data has the variance of error is smaller. Overall, the absolute percentage error is around 10 %. It is promising method for drought prediction.

## 6. Discussion

(7) We were succeeded in improve for the model's evaluation. The base model was set up. In the next step, we will improve the accuracy of the model by using optimized hyperparameters, which were proposed in above. Note that this result is only for first zone. After getting better model, we will analyze the rest of 268 other zones to predict extreme drought trends for South Korea.

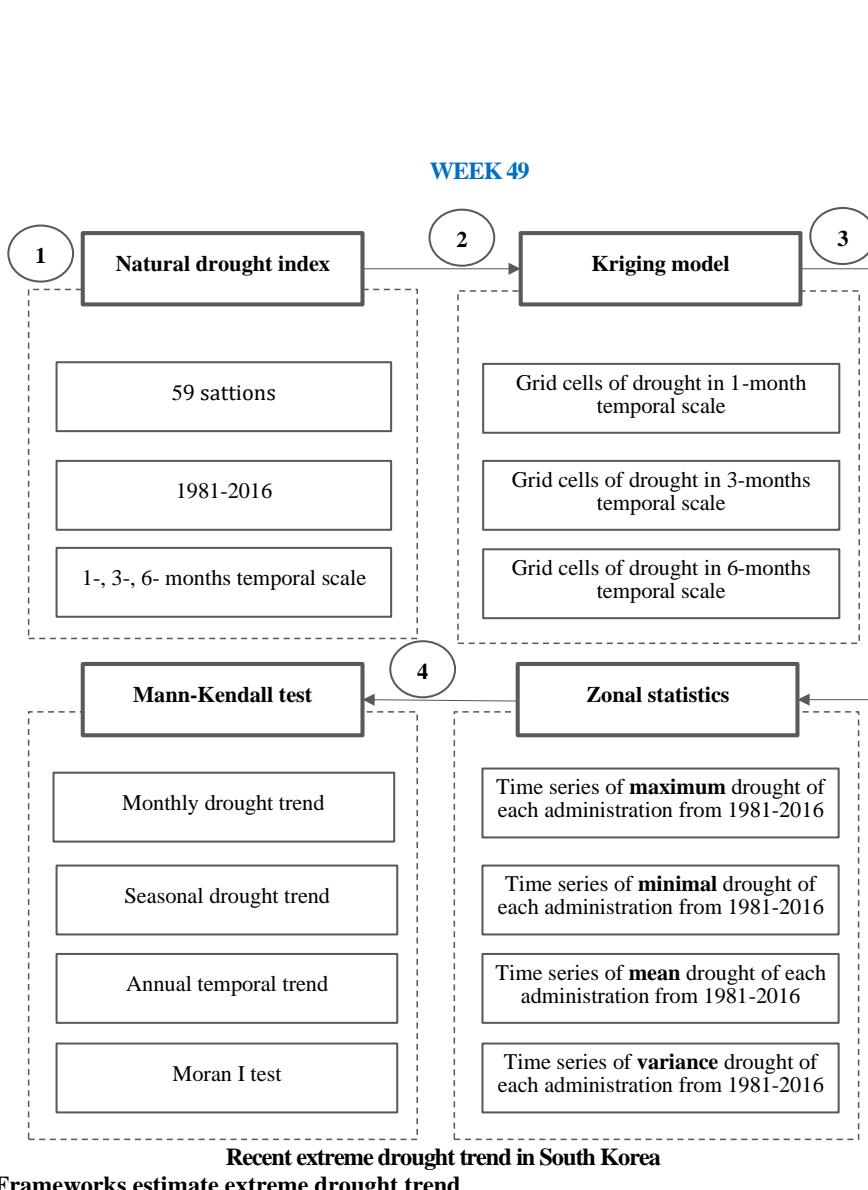
(8) The optimize machine learning is the massive subject. We do not expect understand all of them. In our case, we proposed try to understand the concept, and how to apply in hydrological fields.

## References

- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *The Journal of Machine Learning Research*, 13(1), 281-305.

- Jones, D. R., Schonlau, M., & Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4), 455-492.
- Joseph, R. (2018). Grid Search for model tuning. In.
- Moćkus, J. (1975). *On Bayesian methods for seeking the extremum*. Paper presented at the Optimization techniques IFIP technical conference.

No meeting  
We just follow our planning



#### 7. Frameworks estimate extreme drought trend

The previous aim of our study is estimating the drought trend under multiple time and spatial scale. The Geostatistics, Zonal statistic and non-parametric trend test was utilized to detect spatial and

*Fig 160. Flowchart for extreme drought trend assessment using statistics.*

temporal extreme drought trend. This framework is suitable when the data are assumed independent and randomly ordered (Hamed & Ramachandra Rao, 1998).

The other methods use Machine learning (ML) with the advantages can consider the non-linearity of data and predict the actual values instead of trends. Fig 2 presents a flowchart of framework for extreme drought forecasting. Results of statical framework were reported in the Smart water conference 2020. This report presents the outcome of the ML model.

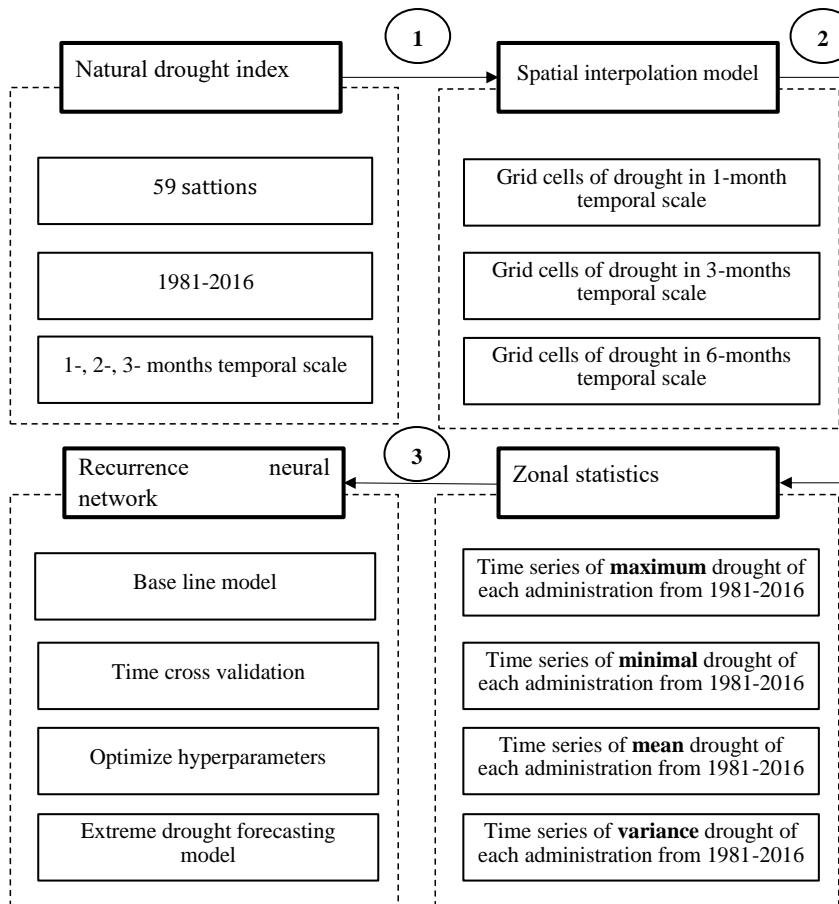


Fig 161. Flowchart for extreme drought trend assessment using machine learning

## 8. Machine Learning model

### 2.9. Baseline model

Stack long short-term memory, one type of RNN was utilized to predict 269 zonal time series. The sequential model includes 1 input layer, 2 hidden layers and 1 output layer with total 511,516 parameters. Results from the base line model were presented in Fig 3.

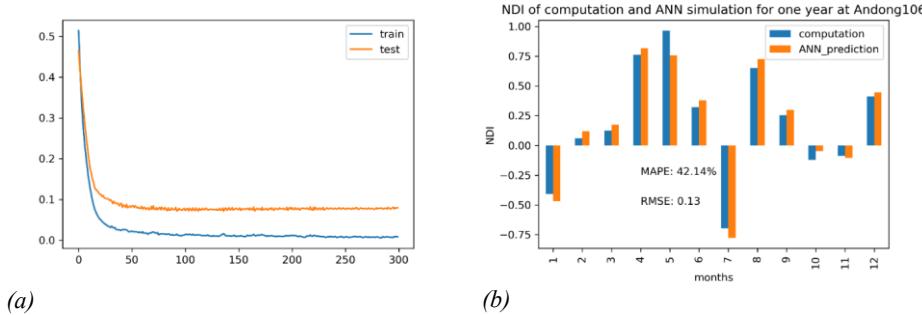
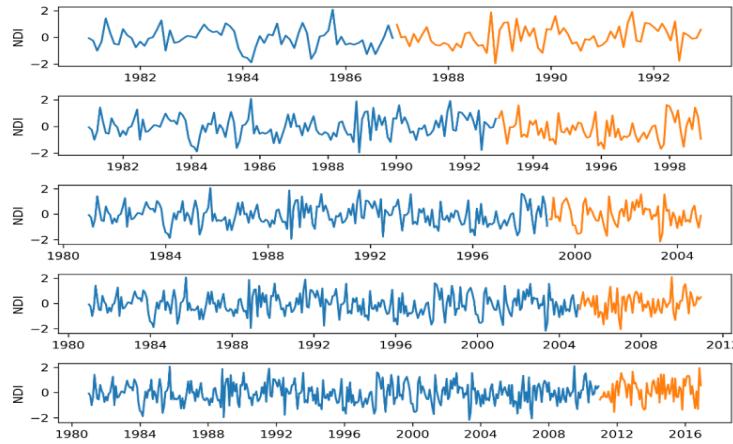


Fig 162. Training/testing processing (a) and (b) initial result at Andong106 zone.

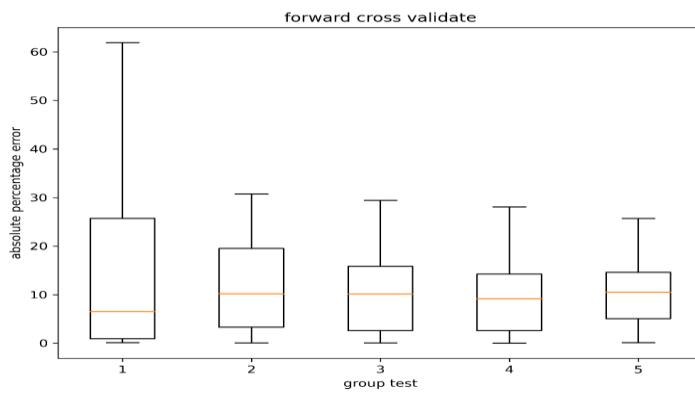
The testing loss is higher than testing loss mean that model is overfitting. It should be checking again.

### 2.10. Time cross-validation

Data was split into 5 cross validation sample (CV) in temporal order. Results of 5 split data were presented on the Fig 4. Each sub-sample has various lengths. The last sub-sample has the highest ratio of training over testing. It has the lowest variance of the absolute percentage error is under 30%. While the first sub-sample has the lowest ratio of training/testing has the highest the absolute percentage error is over 60%.



(a)



(b)

Fig 163. Input cross validations (a), and results (b)

## 2.11. Optimize hyperparameters

Despite of great model, machine learning models are usually characterized by a large set of hyperparameters. They define the network's topology, computational power and so on. As a result, the hyperparameters need to be configured to harness the network's functionality. So, optimization of the hyperparameters' configuration appears to be quite a challenging task, as it differs,

depending on the task being executed, the dataset being used etc., rendering each situation unique (Gorgolis et al., 2019). In this study, we optimize a hyperparameters space of number neuron of hidden layers, batch size, activation function, learning rate and loss functions. The Bayesian, minimal Gaussian progress using efficient improve function (Mockus et al., 1978) was chosen for optimization. Table1 presents space search of hyperparameters.

*Table 22. Space and optimized hyperparameters.*

No.	Name	Space	Optimized
1	Number of neurons	112-187	150
2	Number of epochs	600-1000	800
3	Number of batch size	27-45	36
4	Activation function	Sigmoid, Linear, Tanh, Relu	Linear
5	Learning rate	$2.25 - 3.75 \times 10^{-4}$	0.0003
6	Loss function	mae, mse, msle, mape	msle

The results of optimize hyperparameter are presented in Fig 5. The blue-yellow scale is corresponding to the bad-good optimization. We considered the hyperparameters in a pair. For instance, in the top, number of epoch and number of neurons is normal. It is not good because it is not located in the yellow area. Otherwise, activation function is optimized because it's located in the lower area with yellow color.

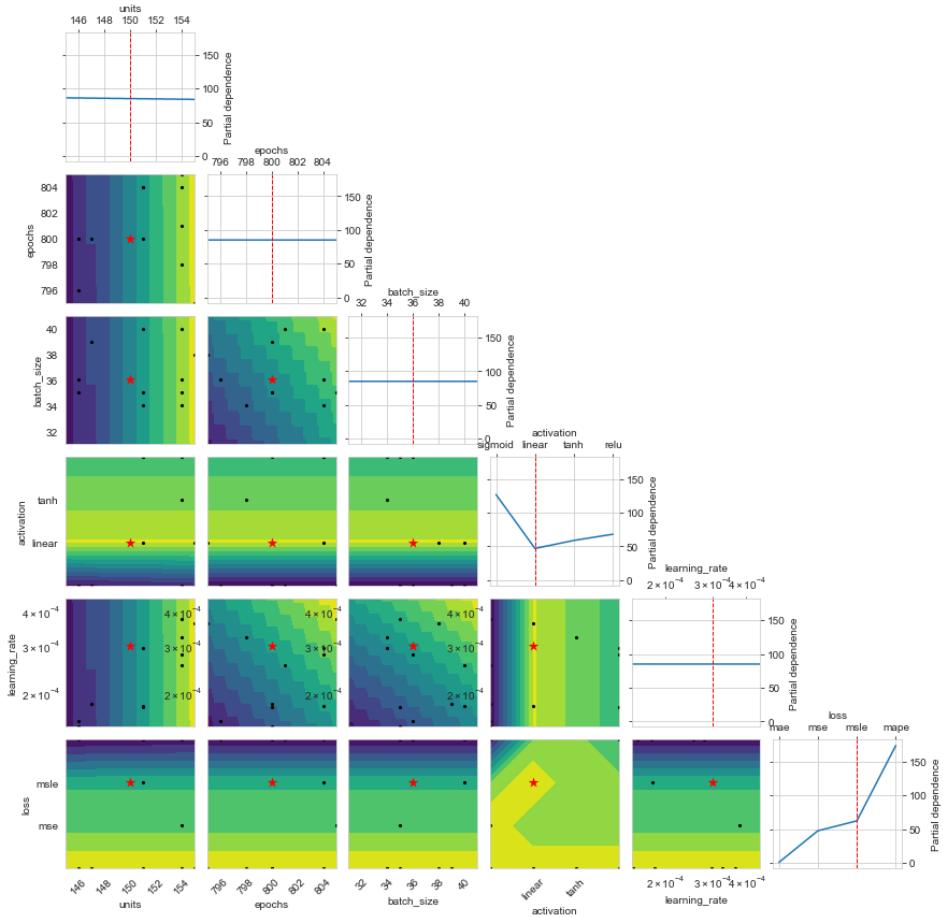


Fig 5 present results of optimizes hyperparameters.

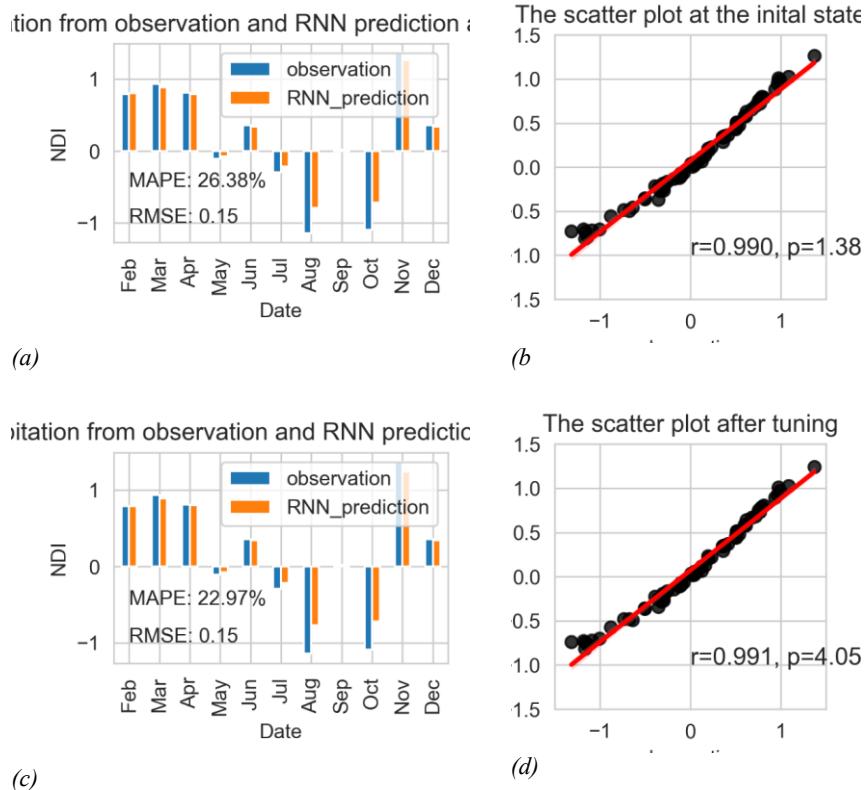


Fig 164. Estimate results of the model before (a, b), and after tuning (c, d).

The model has an improvement with mean absolute percentage error reduce from 26.38 % to 22.97 %. The Pearson correlation of prediction and computed is increased from 0.990 to 0.991 at the confidence level  $> 99\%$ . It is noted that the values of  $p$  in the Pearson correlation test are very small. They are nearly zero. Because the limited space of print, they exponential of  $p$ :  $e^{-60}$ ,  $e^{-63}$  were not printed.

## 9. Discussion

(9) We propose separating our study. One is using statistical method with Mann-Kendal (Fig1) and the other is using Machine learning (Fig 2). By this way, we can analysis deeper for each

method. For instance, we can compare several MK methods with decompose model. In machine learning we can estimate impacts of auto correlation of multiple timeseries.

(10) Optimize hyperparameters is interesting topic. But it requires a huge computer's resources. In our cases, using personal computer (CPU i5-75000 3.40 GHz, 16.0 GB installed memory) take 40 minutes for one simulation. We have 269 zones; it requires more than a week for complete on a scheme. It is difficult because the optimize is hard to be achieved in one simulation. Therefore, we have some options: (I) finding the new fast optimization algorithms; (ii) improve hardware or using another server. Our server was not built for deep learning. It has not enough GPU; (iii) limited search space such as: only optimize number of hidden layer and nodes (Stathakis, 2009), or activation function (Farzad et al., 2019); iv) ignore or using experimental methods to guess the suitable hyperparameters. We strongly believe that optimize subject can be explored in the futures.

(11) Overall, we completed 70% of two studies: estimate extreme drought trend using statistic and forecast extreme drought using machine learning. The study about improve spatial drought coverage using NDI, satellite data and ANN is also 60 % completed.

(12) In the next steps, we propose reviewing all current studies to complete the priority.

## References

- Farzad, A., Mashayekhi, H., & Hassanpour, H. (2019). A comparative performance analysis of different activation functions in LSTM networks for classification. *Neural Computing and Applications*, 31(7), 2507-2521. doi:10.1007/s00521-017-3210-6
- Gorgolis, N., Hatzilygeroudis, I., Istenes, Z., & Gynne, L. G. (2019, 15-17 July 2019). *Hyperparameter Optimization of LSTM Network Models through Genetic Algorithm*. Paper presented at the 2019 10th International Conference on Information, Intelligence, Systems and Applications (IISA).
- Hamed, K. H., & Ramachandra Rao, A. (1998). A modified Mann-Kendall trend test for autocorrelated data. *Journal of Hydrology*, 204(1), 182-196. doi:[https://doi.org/10.1016/S0022-1694\(97\)00125-X](https://doi.org/10.1016/S0022-1694(97)00125-X)
- Mockus, J., Tiesis, V., & Zilinskas, A. (1978). The application of Bayesian methods for seeking the extremum. *Towards global optimization*, 2(117-129), 2.
- Stathakis, D. (2009). How many hidden layers and nodes? *International Journal of Remote Sensing*, 30(8), 2133-2147.

Some comments from Jae-Yeong:

- Validate and testing what is difference?
- Reason to choose activation functions?
- Paper present about time cross validation.
- Explain about multiple timeseries prediction
- Window of LSTM

## WEEK 50

### Recent extreme drought trend in South Korea

#### 1. Frameworks estimate extreme drought trend using machine learning

The other methods use Machine learning (ML) with the advantages can consider the non-linearity of data and predict the actual values instead of trends. Fig 2 presents a flowchart of framework for extreme drought forecasting. Results of statical framework were reported in the Smart water conference 2020. This report presents the outcome of the ML model.

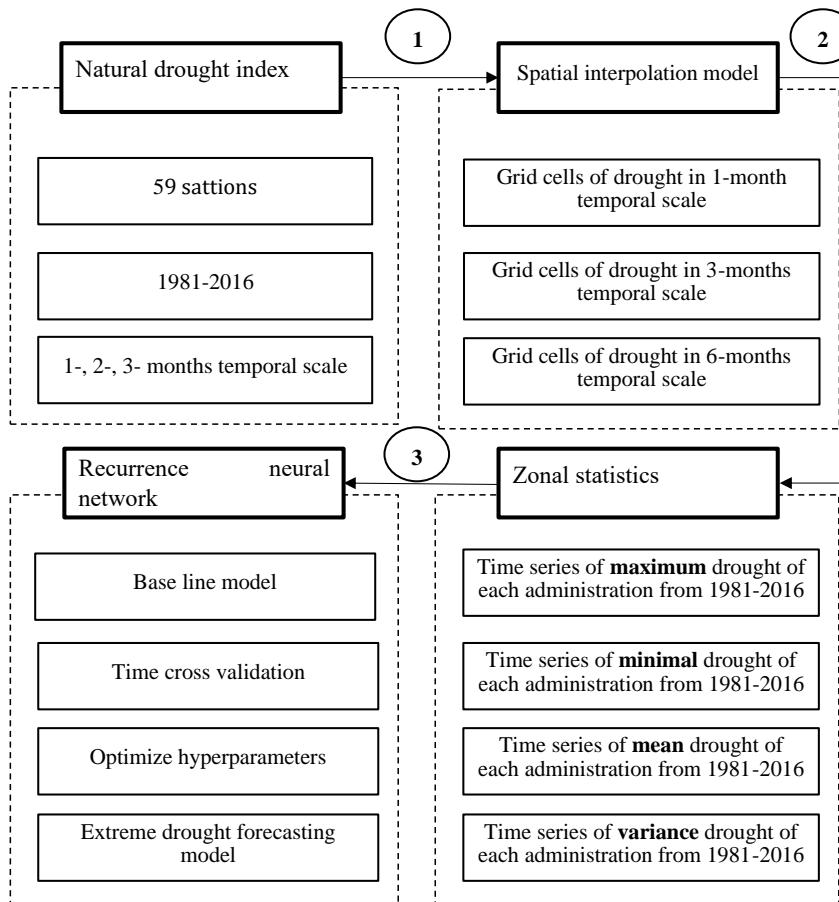


Fig 165. Flowchart for extreme drought trend assessment using machine learning

## **2. Machine Learning model**

### *2.1. Training/testing processing*

Explain input data of testing and processing

### *2.2. Time cross-validation*

Explain time series cross-validation

### *2.3. Optimize hyperparameters*

Give the reason why we chose the initial hyperparameters

- Activate functions
- Loss functions
- Optimizer

## **3. Testing other models**

To choose the suitable model, we compare the typical LSTM model (M0) with 4 models: Encoder-Decoder-LSTM using univariate input (M1), Encoder-Decoder-LSTM using multivariate input (M2), CNN-LSTM-Encoder-Decoder (M3), Conv-LSTM encoder-decoder (M4). These models were estimated that is better than typical LSTM. For instance, Kao et al. (2020) was succeed in using M1 to forecast flooding for The Shihmen Reservoir catchment in Taiwan. Results show that M1 translates and links the rainfall sequence with the runoff sequence can improve the reliability of flood forecasting and increase the interpretability of model internals.

In our study, the outcome of models is presented from Fig 2 to Fig 6. Results show that all models (M1- M4) was outperformed compare with typical LSTM model (M0). M0 give the prediction with the highest error, MSE equal 0.699. M3 gives the best prediction with lowest error MSE values at 0.625.

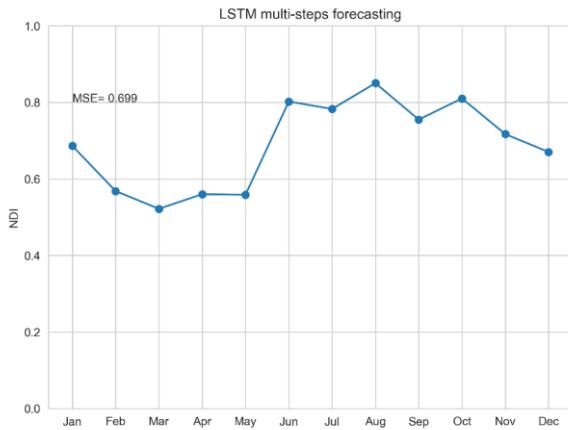


Fig 166. The 12 months NDI predictions of model M0.

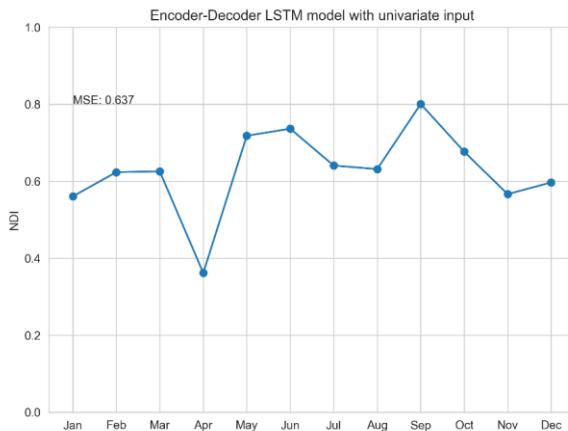


Fig 167. The 12 months NDI predictions of model M1.

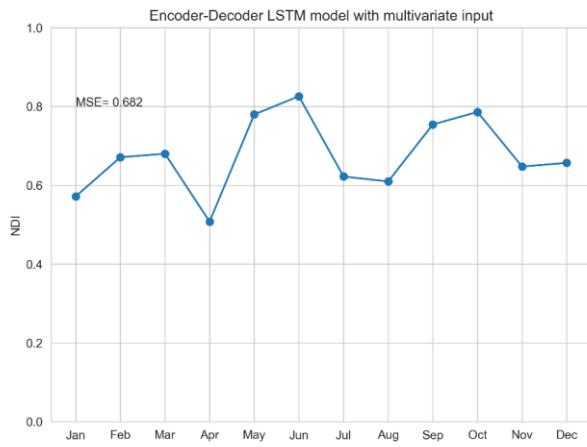


Fig 168. The 12 months NDI predictions of model M2.

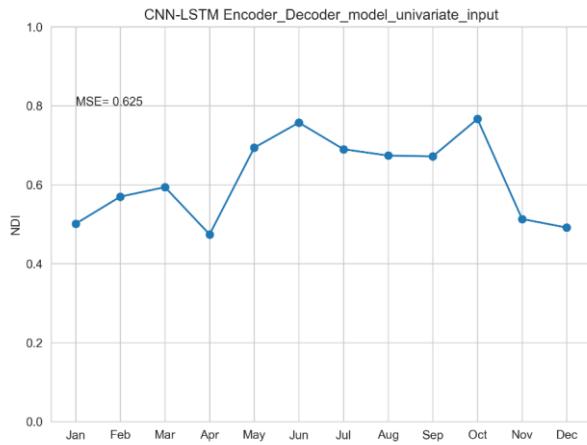


Fig 169. The 12 months NDI predictions of model M3.

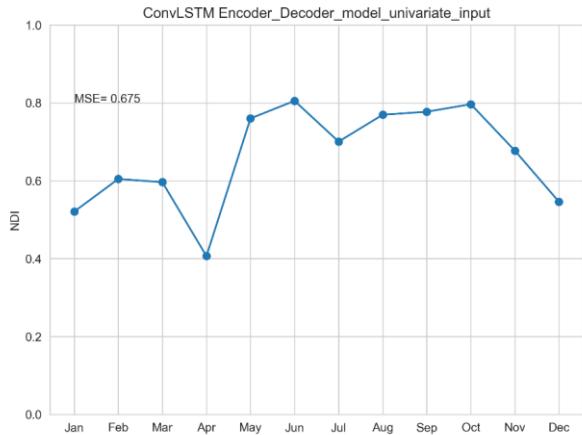


Fig 170. The 12 months NDI predictions of model M4.

#### 4. Discussion

- (1) Reduce the input hyperparameters by removing the unsuitable ones can improve the computing times.
- (2) Cross-validation should be careful selections because of the effective ways to the accuracy of model.
- (3) Multiple time steps prediction was obtained in our study.
- (4) The initial results were obtained. However, we proposed to examine deeply each parts of machine learning which were presented above. It is so needed to understand why results are different from models. Therefore, we would like to exploit these issues in next steps.

#### References

- Kao, I. F., Zhou, Y., Chang, L.-C., & Chang, F.-J. (2020). Exploring a Long Short-Term Memory based Encoder-Decoder framework for multi-step-ahead flood forecasting. *Journal of Hydrology*, 583, 124631. doi:<https://doi.org/10.1016/j.jhydrol.2020.124631>

## WEEK 51

### Basic concept of CNN-LSTM -Encode-decode model

#### 1. Encode-decode model

The encoder-decoder architecture for recurrent neural networks is proving to be powerful on a host of sequence-to-sequence prediction problems in the field of natural language processing such as machine translation and caption generation. Attention is a mechanism that addresses a limitation of the encoder-decoder architecture on long sequences, and that in general speeds up the learning and lifts the skill of the model no sequence-to-sequence prediction problems. In Fig 1 demonstrates a typical Encoder-decoder sequence to sequence model.

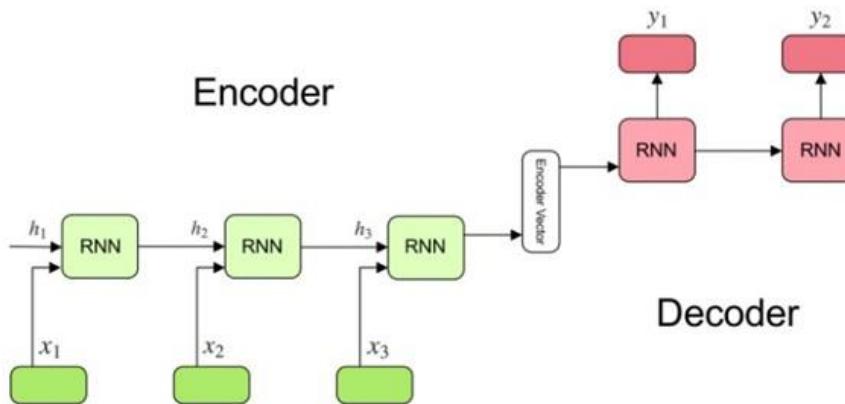


Fig 171. Encoder-decoder sequence to sequence model.

The model consists of 3 parts: encoder, intermediate (encoder) vector and decoder.

#### Encoder

A stack of several recurrent units where each accepts a single element of the input sequence, collects information for that element and propagates it forward. The hidden states  $h_i$  are computed follow the equation:

$$h_t = f(W^{(hh)}h_{t-1} + W^{(hx)}x_t) \quad (1)$$

It is used to apply the appropriate weights to the previously hidden states  $h_{t-1}$  and input vector  $x_t$ .

#### Encoder vector

This is the final hidden state produced from the encoder part of the model. This vector is used to encapsulate the information for all input elements to help the decoder make accurate predictions. It acts as the initial hidden state of the decoder part of the model.

## |Decoder

A stack of several recurrent units where each predicts an output  $y_t$  at a time step t. Each recurrent unit accepts a hidden state from previous unit and produces an output as well as its own hidden state. Any hidden state  $h_i$  is computed follow equation:

$$h_t = f(W^{(hh)} h_{t-1}) \quad (2)$$

The output  $y_t$  at time step t is computed using the equation:

$$y_t = \text{softmax}(W^S h_t) \quad (3)$$

The output was calculated by using the hidden state at the current time step and respective weight  $W^S$ .

The power of this model lies in the fact that it can map sequences of different lengths to each other. The inputs and outputs are not correlated, and their lengths can differ. This opens a whole new range of problems which can be solved using such architecture.

The encoder-decoder can be applied bot for CNN and RNN follow the steps in Fig 2.

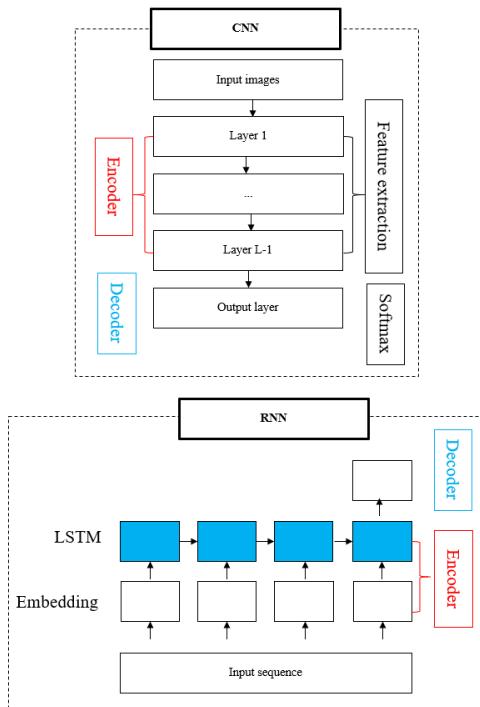


Fig 172. Typical encoder-decoder with CNN and RNN

## 2. CNN-LSTM

The CNN-LSTM has some advantages compared with other methods. Compare CNN-LSTM to Seasonal Autoregressive Integrated Moving Average (SARIMA) model or univariate LSTM, results figured out that CNN-LSTM is outperform in predicting temperature (Kreuzer et al. 2020). The MASE of CNN-LSTM is 0.93, LSTM is 0.99, SARIMA is 1.05.

The CNN-LSTM combines the advantages of CNN and LSTM. The  $m \times c$  variables at  $n$  location in the past  $t$  hours were taken as input data. The input dimension was the same as the CNN, which was  $t \times m \times n$ . The input convolution, pooling layers of CNN were used to extract the features of the input data. The obtained features were flattened into the 1-D array as the time sequence of sequential input data of LSTM. Finally, the prediction values at  $n$  locations in the next time  $t$  were obtained through the fully connection and output layers. The structure of CNN-LSTM was presented in Fig 3.

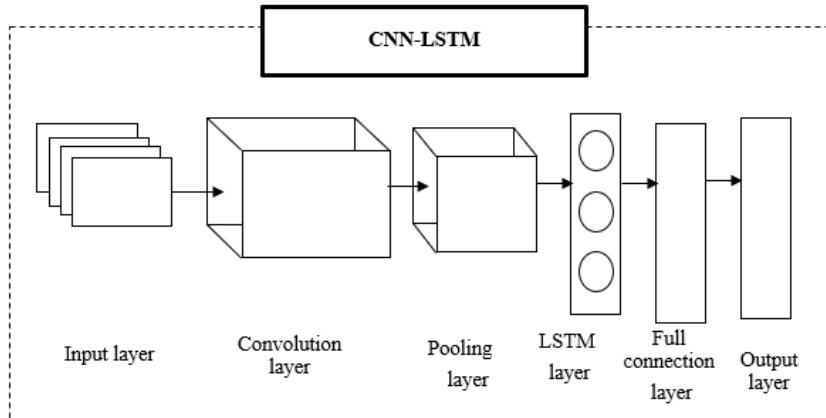


Fig 173. The structure of CNN-LSTM.

In this study, we combined encoder-decoder and CNN-LSTM to predict extreme drought trends.

## 3. Discussion

- (1) The advantages of CNN-LSTM compare to typical LSTM that it cannot only tackle the problem of long time dependence with RNN, but also extract more abundant features with

CNN (Zhang et al. 2018). CNN layers were used to extract spatial features and LSTM was utilized for modeling temporal information (Khan et al. 2020). While The key benefits of the encoder-decoder that approach are the ability to train a single end-to-end model directly on source and target, and the ability to handle variable length input and output sequences.

(2) In the next steps, we propose to restructure of our study. Then we select the most important results to presents in the article. They will be revised if it is needed. Begin writing the initial draft of study “A framework for extreme drought prediction using encoder-decoder-CNN-LSTM”.

## References

- Khan ZA, Hussain T, Ullah A, et al (2020) Towards efficient electricity forecasting in residential and commercial buildings: A novel hybrid CNN with a LSTM-AE based framework. Sensors (Switzerland) 20:1–16. doi: 10.3390/s20051399
- Kreuzer D, Munz M, Schlüter S (2020) Short-term temperature forecasts using a convolutional neural network — An application to different weather stations in Germany. Mach Learn with Appl 2:100007. doi: 10.1016/j.mlwa.2020.100007
- Zhang X, Chen F, Huang R (2018) A combination of RNN and CNN for attention-based relation classification. Procedia Comput Sci 131:911–917. doi: 10.1016/j.procs.2018.04.221

2021

## WEEK 1

### A framework predicts extreme drought using encoder-decoder CNN-LSTM model

#### Motivations:

- Drought is one of the most severe natural disasters. Due to its complexity with diverse origins and occurrence at different temporal and spatial scales, drought prediction has presented a major challenge to climatologists and hydrologists as well as decision and policy makers.
- Statistical methods are usually utilized to predict drought. The statistical prediction method uses empirical relationships of historical records, taking different influencing factors as predictors. Time series model, regression model, Markov chain model, the conditional probability model is popular methods in drought prediction.
- The main time series modeling technique is the Autoregressive Integrated Moving Average (ARIMA) framework. The ARIMA is suitable for predicting several drought indices such as SPI, PDSI. The main limitation of this model is that it assumes a linear relationship between the predictand and predictors, and thus, it generally falls short in capturing nonlinear characteristics. Linear regression is a traditional method for statistical prediction in hydrology and climatology.
- In contrast to the time series model that predicts drought severity based solely on the persistence of certain drought indicators, the regression model seeks to establish the relationship between predictand and other variables that may contribute to the predictive information of drought developments. Logistic regression, multivariate regression are the other types of the regression model.
- In many cases, drought condition is classified into different states (e.g., two states of wet and dry) or categories with a specific threshold. The problem of drought prediction in this case can be formulated as the transition from the wet or normal state to the dry state (or the other way around) and thus the transition probability must be modeled. This can be handled by the Markov Chain (MC) model based on the stochastic process with a countable state space, in which future states are assumed to depend only on the current state.
- The joint distribution has been commonly used to characterize the joint behavior of multiple drought variables, from which the conditional distribution can be constructed to model the predictand conditioned on predictors. The advantage of the conditional probability model in drought prediction is that nonlinear dependences between the predictand and predictors can be modeled, and probabilistic prediction can be achieved from the conditional distributions. The temporal dependence (different time lags) and cross-variable dependence of different locations (spatial dependence) can be modeled with the function F for drought prediction. The main disadvantage of this type of model is that when a relatively large number of predictors are involved, the joint distribution in high dimensions is difficult to build. Though certain models may be used for statistical inference of multiple variables (e.g., multivariate normal distribution), they are generally limited in the dependence modeling in high dimensions. The recently developed vine copula is capable of modeling flexible dependences of multiple drought indicators and influential factors.
- Machine learning method based on the data driven has been applied in various disciplines. In recent, several studies have figured out the positive of machine learning model for drought prediction. The recurrent neural network (RNN) and convolution neural network (CNN) are two of among deep learning branches. RNN is often used to analyze sequential data, while CNN is used to extract features of the images.
- The previous studies were mainly using one type of deep learning algorithms (RNN, CNN or encoder-decoder). The combination of various models were not much utilized in drought prediction. To take the advantages of various deep learning algorithms, we proposed a deep learning model combines CNN-LSTM and encoder-decoder for drought prediction.

#### Objectives:

- The aim of this study is development a framework for drought prediction using deep learning approach.

**Methodology:**

- The methodology for extreme drought prediction is summarized in **Fig 1**. The procedure consists of 4 steps. Step 1 computes NDI at various scales 1-, 3-, 6- months temporal scale. Step 2 interpolated data to build a continuous model using Kriging model. Step 3 determines the minimal, maximal, mean, and extreme drought for 269 administrative districts. Final steps, using multivariate time series drought of administrative districts to predict the extreme drought. The concept of the encoder - decoder and the CNN - LSTM model were presented in **Fig 3** and **Fig 4**. Using the predict extreme drought we derived the futural drought map which are used for support water resource managements and water planning.

**Study area and data:**

- Study area: Study area is South Korea with a total area of 100,032 square km. Data: NDI was computed at the 59 Automatic Synoptic Observation System (ASOS).

**Results:**

- The spatial distribution of extreme drought from Kriging model in September 2015 was presented in **Fig 5**.
- The cross-validation for time series was presented in **Fig 6**.
- The optimal parameters were presented in **Table 1** and **Fig 7**.
- Model evaluation was presented in the **Fig 8**.
- The comparison form various deep learning models for drought prediction was presented in **Fig 9**.
- The spatial predicted drought was presented in **Fig 10** (it likes to **Fig 5** and was not completed).

**Conclusions:**

- Predict the temporal-spatial extreme drought is one of the most important issues in water resource planning and management.
- This study predicts temporal-spatial extreme drought using natural drought index, geo-statistic, and deep learning approach.
- Results show that model is promising method for drought predictions.

## Appendix

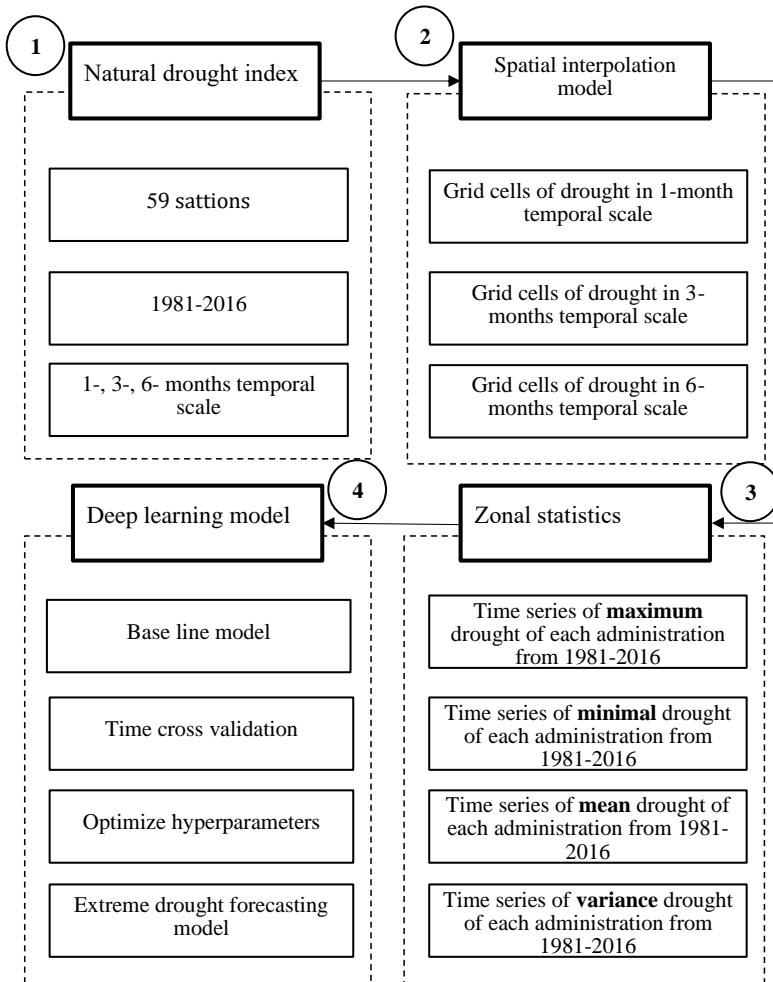


Fig 174. Flowchart for drought prediction using deep learning approach

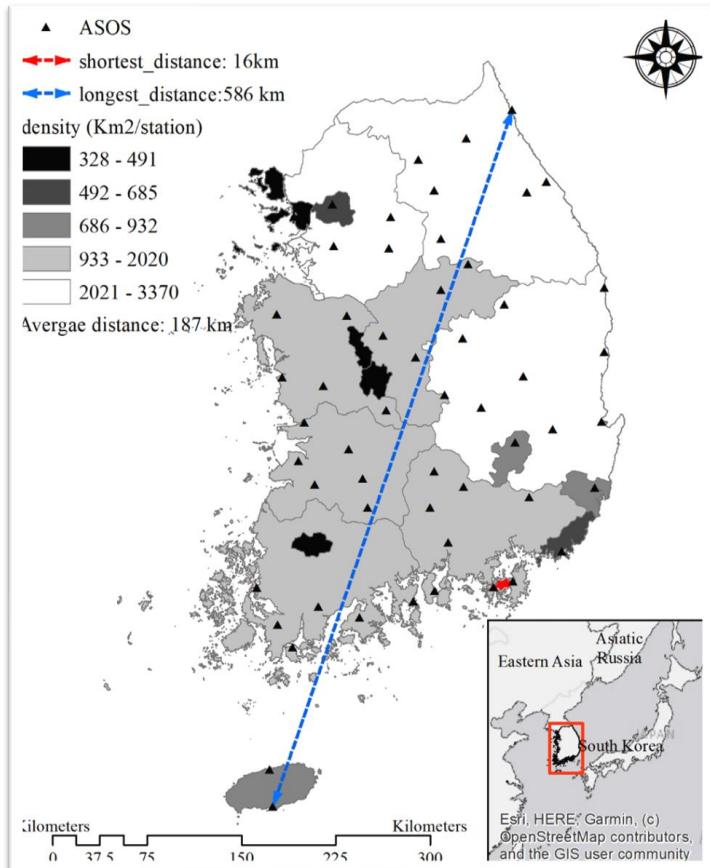


Fig 175. ASOS locations and study area

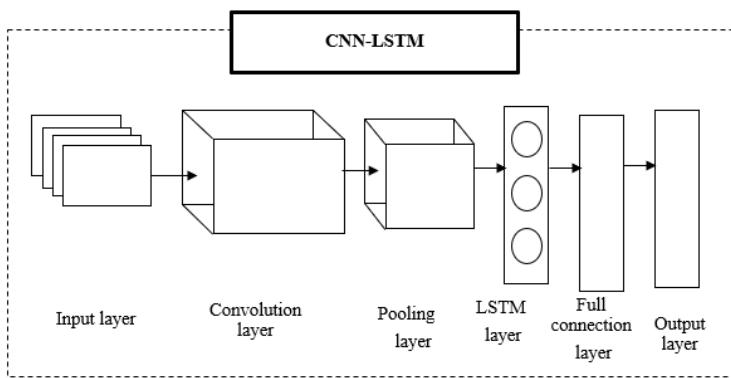


Fig 176. The structure of CNN-LSTM.

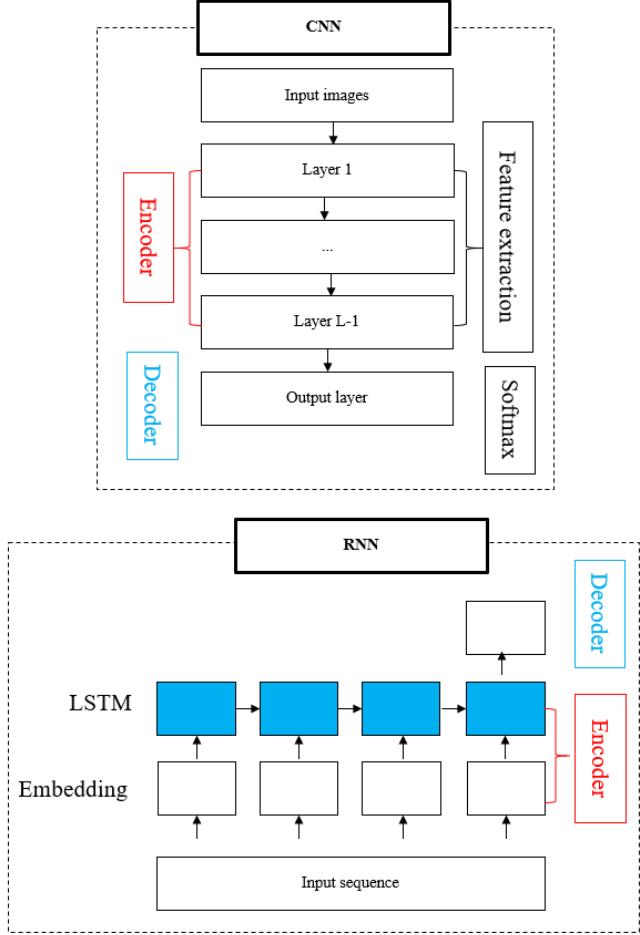


Fig 177. Typical encoder-decoder with CNN and RNN

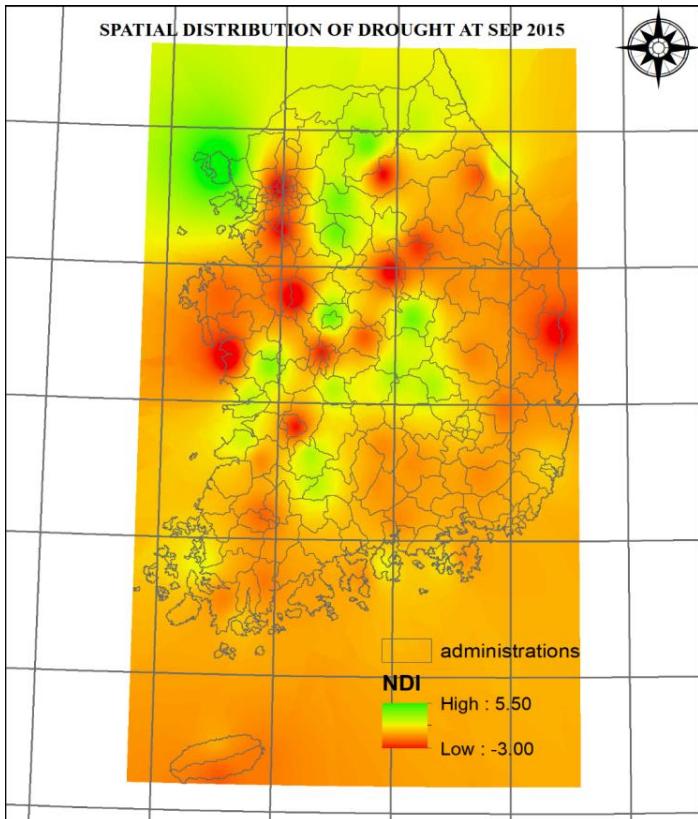


Fig 178. Spatial drought using NDI in September 2015.

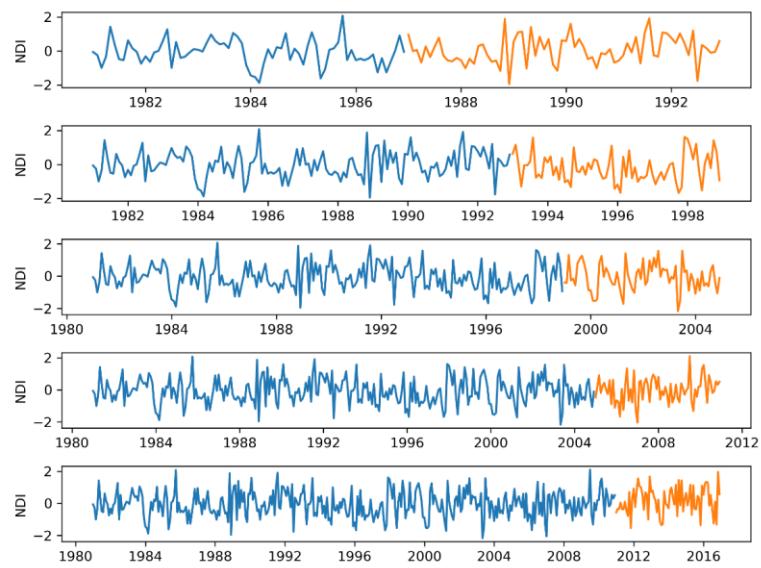


Fig 179. CV with 5 split parts for Ansan75 zone.

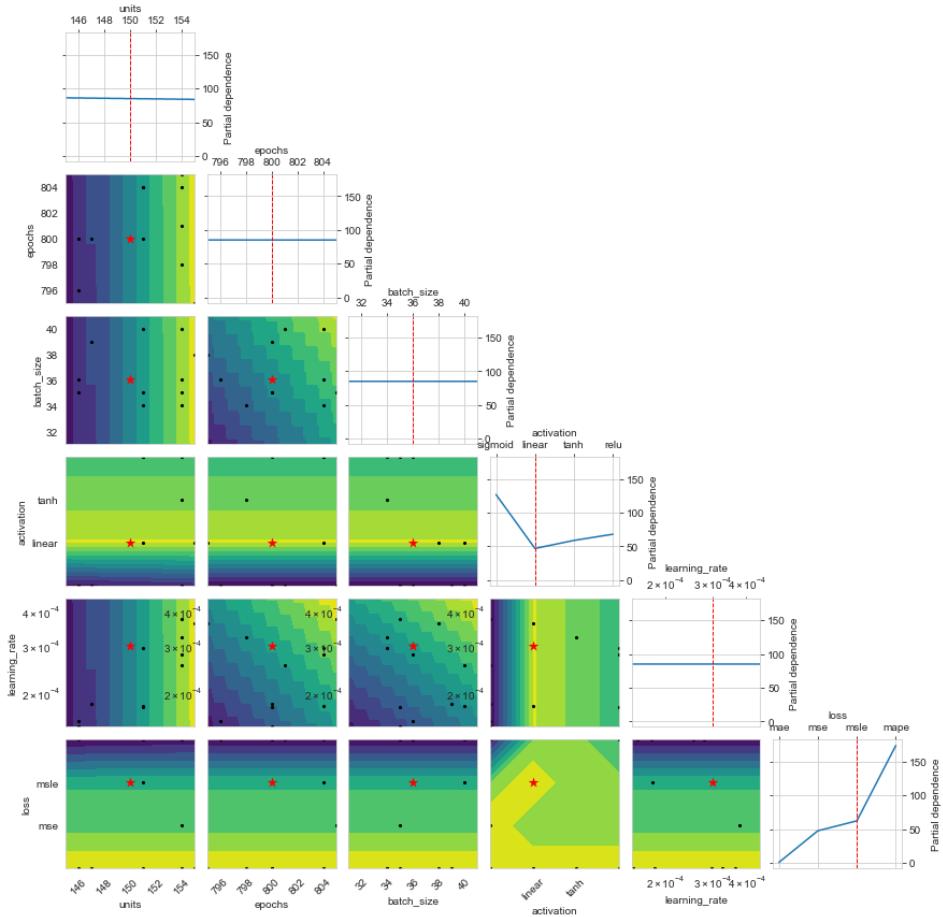
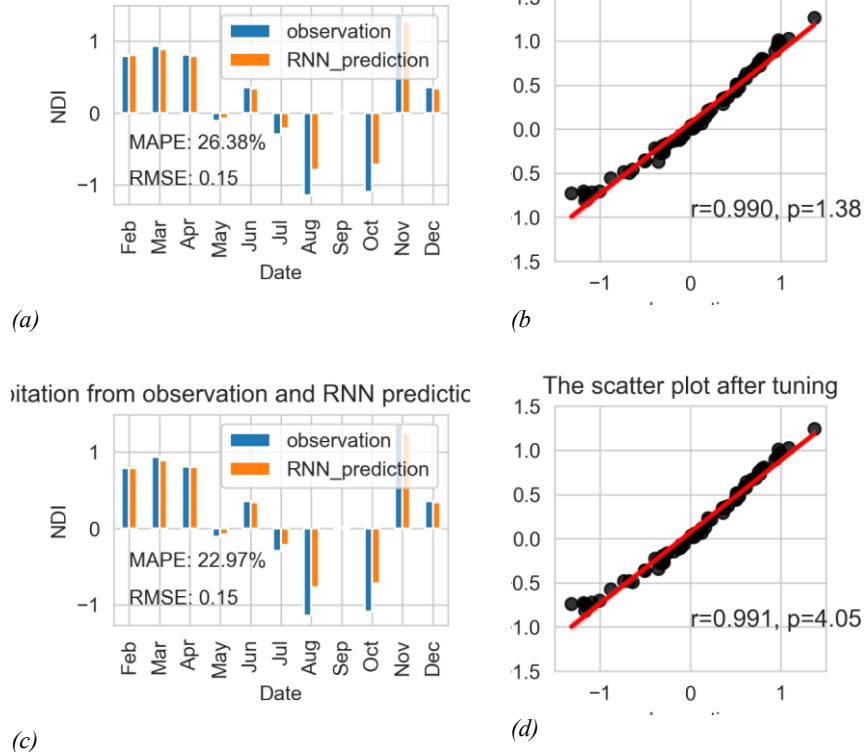
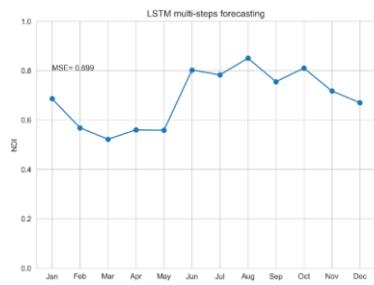


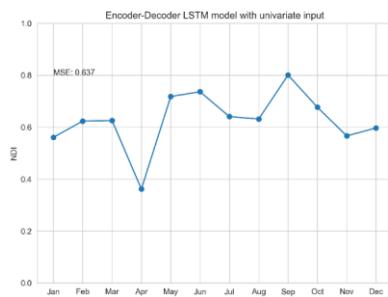
Fig 180. Results of optimizes hyperparameters.

Fig 181. Estimate results of the model before (a, b), and after tuning (c, d).

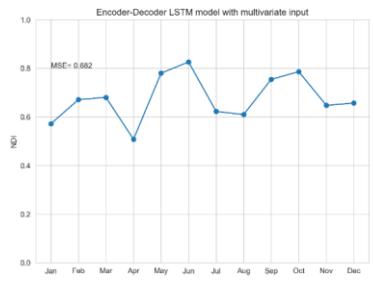




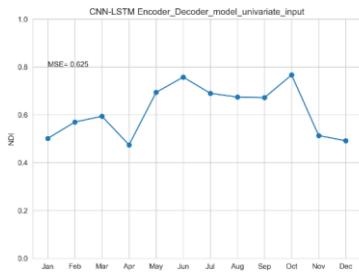
(a)



(b)



(c)



(d)

Fig 182. Drought prediction based on LSTM (a), encoder-decoder LSTM univariate (b), encoder-decoder LSTM multivariate, encoder-decoder CNN-LSTM (c), encoder-decoder Conv-LSTM (d).

**Commented [VQT5]:** This is MAE error. It is not predicted NDI values.

*Table 23. Space and optimized hyperparameters.*

No.	Name	Space	Optimized
1	Number of neurons	112-187	150
2	Number of epochs	600-1000	800
3	Number of batch size	27-45	36
4	Activation function	Sigmoid, Linear, Tanh, Relu	Linear
5	Learning rate	$2.25 - 3.75 \times 10^{-4}$	0.0003
6	Loss function	mae, mse, msle, mape	msle

Profesor comments:

- Make the structure of study for easy to follow
- Every week put the material in it

## WEEK 1

### **Weekly report detail**

Last week, we focused on revising the title and clarifying the study flowchart. The detail of each part was presented, followed by.

#### **1. Revise the title.**

We proposed three initial titles for our study.

- The first title “**Improve extreme drought prediction based on natural drought index, geostatistical and deep learning approaches.**”
- The second title “**Improve extreme drought prediction based on incorporated multiple deep learning models.**”
- The third title “**Can we improve extreme drought prediction based on incorporated multiple deep learning models.**”

The first title consists of the topic, objective, method, results. The objective of the study is improving the accuracy of drought prediction. Geostatistical and deep learning are two main methodologies. The reader can predict the content of our study. It could include a natural drought index, one geostatistical model and deep learning model. They are interesting because of the message. The novelty of incorporates various deep learnings and Geostatistics to forecast extreme drought. This title was made follow the descriptive type. The length of title is suitable (5-15 words) follow the guidance online of the University of Southern California. It has also keywords: natural drought index, geostatistical, deep learning that readers can search easily.

The second title emphasizes the uniqueness of our study is a combination of various deep learning for drought prediction. It attracted the readers who are interested in using deep learning in hydrology fields. It provides the general concepts of using simultaneously various models. In statistical field, using the multiple ensemble models to forecast has been paid in attention from researchers.

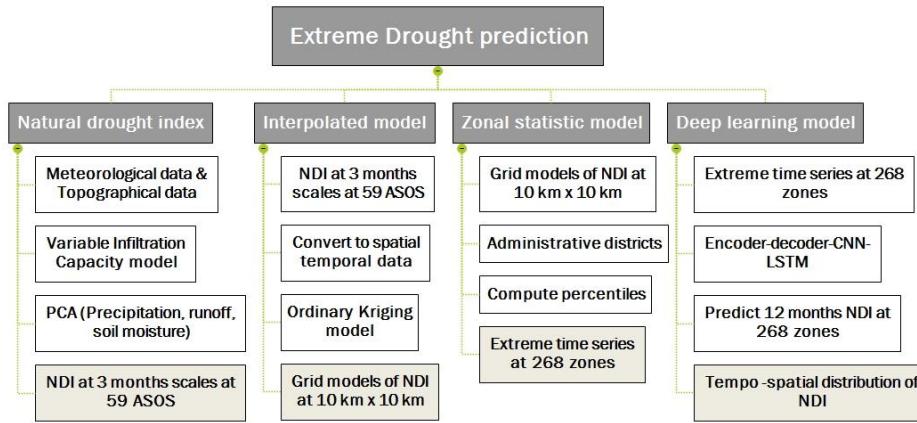
The third title was created follow the interrogative type. In this type of title, the scientific question was stated to catch readers' attention. However, the question type is rarely used in the publication. For instance, based on statistic, Anthony (2001) found only two titles in the form of questions from a sample of 600 articles in computer science.

#### **2. Revise the flowchart of the framework.**

The input and output were added in the flowchart of study (Fig 1) to make easier for readers. The inputs locate at rectangle without filling colors, and the output lays on the rectangle with filled color at the end of sections. The framework consists of 4 sections (natural drought index, interpolate model, zonal statistic, and deep learning model). Section 1 (Natural drought index) used VIC model, PCA as the inputs. Monthly NDI with 3-months scales at 59 ASOS from 1981-

2016 are output. In section 2 (the interpolated model), we convert NDI at 59 ASOS to spatial-temporal data by adding coordinates. Then, using Kriging model to create a Grid model of NDI at spatial resolution 10km (10 km). This grid model and administrative districts are the input of section 3 (zonal statistic model). The percentile at 10<sup>th</sup> was selected as extreme level for 268 zones.

In section 4 (Deep learning model), 268 zones were used as the inputs of Encoder-decoder-CNN-LSTM model. Using multiple steps, we predict 12 months NDI at 268 zones. Final map of monthly NDI for the next annual period is our expectation.



*Fig 183. Flowchart of our study.*

#### 2.4. Natural drought index

Kim et al. (2016) proposed a method to calculate the NDI using the VIC model and principal component analysis (PCA). PCA is a method that produces new variables through a linear combination of the original variables. High-dimensional data can be reduced to low-dimensional data. Precipitation, runoff, and soil moisture are the three components that were used in the PCA to compute the NDI. Runoff and soil moisture data were extracted from the VIC model. The infiltration, evapotranspiration and other hydrological processes were included to simulate runoff and soil moisture. These data have a spatial resolution of 1/8 degrees and a temporal scale of 24 hours. Daily precipitation was measured by ASOS at a spatial resolution of 12 km. NDI was analyzed based on the spatial resolution of the 59 ASOS stations and a monthly temporal scale. Precipitation data were matched to the closest runoff and soil moisture grid cell using the nearest-neighbor method. The data were accumulated to match the monthly temporal scale used to compute the NDI. The seasonal influences of variables on drought were also considered; the months were treated separately. An extreme drought event was identified when  $NDI \leq -2$ , which

indicated a probability value less than or equal to 0.02 and corresponded to a 50-year return period. The detailed information on the classification of drought based on the NDI can be found from Kim et al. (2016).

### 2.5. Interpolated model

The geostatistical Kriging model is chosen for creating the continuous surface drought. The spatial configuration of the sampling points around the predicted location. It represents the family of generalized least square regression-based interpolation methods. It aims to minimize mean squared error in prediction.

$$\hat{Z}(x_0) = \sum_{i=1}^N w_i [Z(x_i) - \mu(x_0)] + \mu \quad (45)$$

Here,  $\hat{Z}(x_0)$  is the predicted parameter value at point  $x_0$ ,  $\mu$  is the constant mean value over the region of interest. The OK assumes the stationarity of the first moment of the prediction parameter, that is  $E\{Z(x_i)\} = E\{Z(x_0)\} = \mu = \mu(x_0)$ ,  $\mu$  is unknown. The  $w_i$  is measured from the experimental semivariogram.

$$\gamma(h) = \frac{\sum_{i=1}^N [Z(x_i) - Z(x_i + h)]^2}{2M} \quad (46)$$

Here,  $\gamma(h)$  represents the semi variance at lag interval  $h$ ,  $Z(x_i)$  is measured parameter value at point  $x_i$ ,  $Z(x_i + h)$  is measured parameter value at sampled location which is  $h$  lag distance apart from  $x_i$ ,  $M$  is the total number of pairs of the interpolating points that are  $h$  distance lag apart. Detail of semantic Kriging for spatial-temporal prediction can reference in studies (Bhattacharjee et al., 2019).

### 2.6. Zonal statistic model

Zonal statistic methods summarize and aggregate the raster values intersecting a vector geometry. For instance, zonal statistics provide the mean precipitation or maximum elevation of an administrative unit. Additionally, functions are provided to query a raster at a point and get an interpolated value rather than the simple nearest pixel. The values within the zone were assumed normalized. The percentile was computed based on mean and standard deviation. The extreme value of zone use percentile method Q1 (Hyndman and Fan 1996) for query. The extreme drought

values are determined when NDI equal or below to 10<sup>th</sup> percentiles or corresponds to  $z_{score} = -1.282$ . The mean ( $\mu$ ), standard deviation ( $\sigma$ ),  $z_{score}$  and percentile values of zone ( $x_p$ ) was computed follow equation (3, 4, 5, 6) below:

$$\mu = \frac{1}{N} \sum_i^N x_i \quad (47)$$

$$\sigma = \sqrt{\frac{1}{N} \sum_i^N (x_i - \mu)^2} \quad (48)$$

$$z_{score} = \frac{x - \mu}{\sigma} \quad (49)$$

$$x_p = z_{score} \times \sigma + \mu \quad (50)$$

## 2.7. Deep learning models

The deep learning model uses encoder-decoder model, combine with both LSTM, CNN.

### 1. Encoder-decoder model

The encoder-decoder architecture for recurrent neural networks is proving to be powerful on a host of sequence-to-sequence prediction problems in the field of natural language processing such as machine translation and caption generation. Attention is a mechanism that addresses a limitation of the encoder-decoder architecture on long sequences, and that in general speeds up the learning and lifts the skill of the model no sequence-to-sequence prediction problems. The model consists of 3 parts: encoder, intermediate (encoder) vector and decoder.

Encoder stacks of several recurrent units. Each step accepts a single element of the input sequence, collects information, and propagates to forward. The hidden states  $h_i$  are computed follow the equation:

$$h_t = f(W^{(hh)}h_{t-1} + W^{(hx)}x_t) \quad (51)$$

It is used to apply the appropriate weights to the previously hidden states  $h_{t-1}$  and input vector  $x_t$ .

Encoder vector is the final hidden state produced from the encoder part. This vector is used to encapsulate the information for all input elements to help the decoder make accurate predictions. It acts as the initial hidden state of the decoder part of the model.

Decoder stacks of several recurrent units where each predicts an output  $y_t$  at a time step t. Each recurrent unit accepts a hidden state from previous unit and produces an output as well as its own hidden state. Any hidden state  $h_i$  is computed follow equation:

$$h_t = f(W^{(hh)}h_{t-1}) \quad (52)$$

The output  $y_t$  at time step t is computed using the equation:

$$y_t = \text{softmax}(W^S h_t) \quad (53)$$

The output was calculated by using the hidden state at the current time step and respective weight  $W^S$ .

The power of this model lies in the fact that it can map sequences of different lengths to each other. The inputs and outputs are not correlated, and their lengths can differ. This opens a whole new range of problems which can be solved using such architecture.

## 2. CNN-LSTM model

The CNN-LSTM combines the advantages of CNN and LSTM. The  $m \times c$  variables at n location in the past t hours were taken as input data. The input dimension was the same as the CNN, which was  $t \times m \times n$ . The input convolution, pooling layers of CNN were used to extract the features of the input data. The obtained features were flattened into the 1-D array as the time sequence of sequential input data of LSTM. Finally, the prediction values at n locations in the next time t were obtained through the fully connection and output layers.

## 3. Discussion

(1) The title is one of the most important parts of the article. It is the first reading from readers. On the other word, it is label of sale product. Therefore, we should pay attention. At this time, we finished with making initial title. But it should be adjusted again before submission.

(2) The main parts of the methodology or framework was revised. It provides how we improved extreme drought prediction in theory. But it should be shown in the results. The initial results were

obtained but it is not complete. Therefore, we proposed finishing the results. Otherwise, we can compute the results after complete writing introduction. The introduction part sets the background for the picture that we will hang. It is also one of the most difficult part of writing. Due to lack of experience in writing, I suggest take the times for review some writing books.

## References

- Anthony, L. (2001). Characteristic features of research article titles in computer science. *IEEE Transactions on Professional Communication*, 44(3), 187-194.

## **WEEK 2**

**Improve extreme drought prediction based on incorporated multiple deep learning models.**

Quang-Tuong Vo, Deg-Hyo Bae\*

Department of Civil & Environmental Engineering, Sejong University, 98 Gunja-Dong,  
Gwangjin-Gu, Seoul, 143-747, South Korea

**\*Corresponding author:**

Prof. Deg-Hyo Bae

Department of Civil and Environmental Engineering, Sejong University, 98 Gunja-Dong,  
Gwangjin-Gu, Seoul, 143-747, South Korea

Email: dhbae@sejong.ac.kr

**Abstract**

**Keywords**

#### 4. Introduction

Drought is one of the serious disasters impacts human, environment, and economic (Stahl et al., 2015; Wilhite, 2000). For instance, the total cost damage of drought in the USA from 1980 to 2020 exceeds 1.875 trillion US dollar (NOAA, 2020). In Italia, the total damages to the agriculture cause of drought ranged from 0.55 to 1.75 billion euro (García-León et al., 2021). In addition, the future global socioeconomic risk due to drought has been increasing. Several locations were projected under the highest drought risk condition such as in the Feast and South Asia, Midwestern Europe, eastern US, and coastal area of South America (Y. Liu & Chen, 2021). They showed the drought damages and risk had occurred and required us to have a strategy to expose it. Therefore, forecasting drought is the key task for drought preparedness.

Drought prediction has challenged because it occurred in spatial and temporal dimensions (Hao et al., 2018). Drought does not occur at a point, or constant specific times. Otherwise, drought onset, termination is differing from temporal and spatial factor. It is impacted by the local condition (topology, land used, soil characteristic), and global atmospheric ENSO (Gupta & Jain, 2021). Drought prediction is also challenging due to affected by climate changes, and human activities (Jehanzaib et al., 2020).

To address the future drought, several methodologies such as dynamic model, statistical method, machine learning were used. Dynamic models are methods that are used to forecast based on Climate model, hydrological model. These models simulate physical processes of the atmosphere, ocean, and land. The difference between of models and coarse resolution of climate model prediction, the post-processing technique and multiple technique has been used to enhance the accuracy of prediction. Yao et al. (2020) project standardized precipitation and evapotranspiration index (SPEI) at 3-, 6- and 12-month timescales for the period 2011–2100. Two representative concentration pathway (RCP) scenarios – RCP 4.5 and RCP 8.5 in mainland China was using in the predicted model. Hydrological model based on the physical concepts. It reflected the physical characteristic of the phenomenon. However, the uncertainty of parameters is hard to be obtained. For instance, the uncertainty in hydrology of multiple parameters and multi-GCM prediction (Her et al., 2019).

Due to the random, stochastic of drought, Statistical methods are another approach for prediction. Arnone et al. (2020) predict drought based on the climate season. Based on seasonality authors predict 6 months of upcoming season drought. The accuracy of drought occurred ranged from 1 to 50%, according to the predicted period. The standard precipitation index (SPI) was selected to determine drought. The statistical method usually uses the characteristic of historical events to predict the future. The seasonal, frequency of historical drought is used to predict long-term extreme drought (Zhao et al., 2020).

The hybrid models are cooperated the statistical and dynamic model (Strazzo et al., 2019). For instance, a hybrid model was utilized for predicting drought in China with SPI (Xu et al., 2018). Results showed that the hybrid model can improve the performance of drought forecast by combining statistical and dynamic models using Bayesian model and averaging (BMA) method. The statistical model and hybrid model have some short outcome because of the limitation in theoretical assumption. For example, the regression analysis, one of the most popular methods for

drought prediction show the weak performance in long-lead time prediction. Because it assumes that linearity between of predictor and predictand (Fung et al., 2020).

In a decade, machine learning, and sub fields of machine learning such as deep learning has been applied successfully in various fields (Ferreira et al., 2019; Gao et al., 2019; LeCun et al., 2015; Lee et al., 2020). Application of deep learning in water resources has been reviewed systematically by Shen (2018). Using long short-term memory (LSTM) for drought prediction was proposed by Dikshit et al. (2021). One dimensional neural network (1D CNN) can be used in the time series forecast (Barzegar et al., 2020; Sayeed et al., 2020). The 1D CNN can combine with Long Short-Term Memory model (LSTM) to create a hybrid model Convolutional Neural Network- Long Short-Term Memory (CNN-LSTM) (Barzegar et al., 2020; Kim & Cho, 2019). The result shows the output of CNN-LSTM is better than other modes such as a linear regression model, Random forest regression, Decision tree, and Multilayer perceptron. Autoencoder (AE) is architect is used with CNN (Wu et al., 2020), or LSTM (Heryadi, 2019) for improve weather forecast. The advance of autoencoder is reduce dimension (Hinton & Salakhutdinov, 2006), and handle numerous data (J. N. Liu et al., 2015).

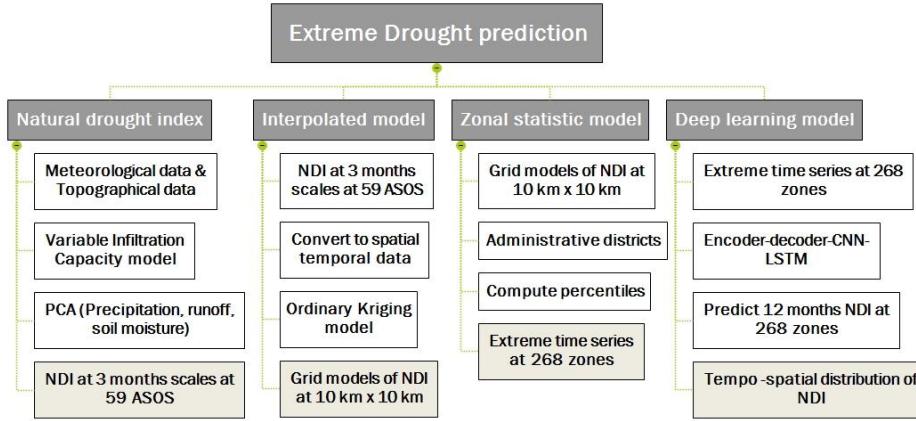
These studies about shows that deep learning stacks multiple hidden layer that can extract more information from the data, Therefore, it can handle with both linearity- and non-linearity (Goodfellow et al., 2016). Although, these deep learning has been succeeded in several cases, the incorporated of multiple models in drought predicted is lack attentions. In this study, we proposed using intercorporate of three model: CNN, LSTM, AE to predict extreme drought. Based on our knowledge, at the current, this is the first study using interoperate of CNN, LSTM and AE for drought projection.

The rest of the article is organized as follows. In section 2, we present material and methods. It includes the structure of CNN, LSTM, AE model. Section 3 present results of our proposed framework. Result of hyper-parameters, model evaluation, and prediction is also found in this section. Some discussion about the advantages, disadvantages, applicable of model is presented in section 4. Finally, the main conclusions about framework were presented in the last section.

## 5. Data and methodology.

The processing of proposed framework was presented in the Fig 1. The inputs are located at rectangle without filling colors, and the output lays on the rectangle with filled color at the end of sections. The framework consists of 4 sections (natural drought index, interpolate model, zonal statistic, and deep learning model). Section 1 (Natural drought index) used VIC model, PCA as the inputs. Monthly NDI with 3-months scales at 59 ASOS from 1981-2016 are output. The study area and statistic distribution of 59 ASOS was presented in Fig 2. In section 2 (the interpolated model), we convert NDI at 59 ASOS to spatial-temporal data by adding coordinators. Then, using Kriging model to create a Grid model of NDI at spatial resolution 10km (10 km. This grid model and administrative districts are the input of section 3 (zonal statistic model). The percentile at 10<sup>th</sup> was selected as extreme level for 268 zones.

In section 4 (Deep learning model), 268 zones were used as the inputs of Encoder-decoder-CNN-LSTM model. Using multiple steps, we predict 12 months NDI at 268 zones. Final map of monthly NDI for the next annual period is our expectation.



*Fig 184. Flowchart of our study.*

## 2.8. Natural drought index

Kim et al. (2016) proposed a method to calculate the NDI using the VIC model and principal component analysis (PCA). PCA is a method that produces new variables through a linear combination of the original variables. High-dimensional data can be reduced to low-dimensional data. Precipitation, runoff, and soil moisture are the three components that were used in the PCA to compute the NDI. Runoff and soil moisture data were extracted from the VIC model. The infiltration, evapotranspiration and other hydrological processes were included to simulate runoff and soil moisture. These data have a spatial resolution of 1/8 degrees and a temporal scale of 24 hours. Daily precipitation was measured by ASOS at a spatial resolution of 12 km. NDI was analyzed based on the spatial resolution of the 59 ASOS stations and a monthly temporal scale. Precipitation data were matched to the closest runoff and soil moisture grid cell using the nearest-neighbor method. The data were accumulated to match the monthly temporal scale used to compute the NDI. The seasonal influences of variables on drought were also considered; the months were treated separately. An extreme drought event was identified when  $NDI \leq -2$ , which indicated a probability value less than or equal to 0.02 and corresponded to a 50-year return period. The detailed information on the classification of drought based on the NDI can be found from Kim et al. (2016).

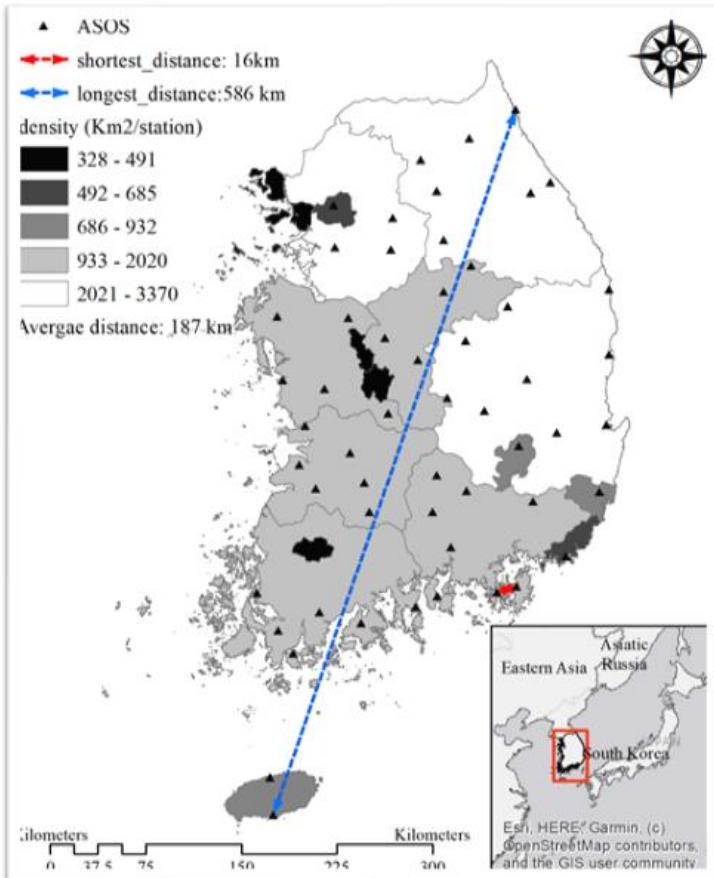


Fig 185. Study area and statistic distribution of NDI at 59 ASOS.

#### 2.9. Interpolated model

The geostatistical Kriging model is chosen for creating the continuous surface drought. The spatial configuration of the sampling points around the predicted location. It represents the family of generalized least square regression-based interpolation methods. It aims to minimize mean squared error in prediction.

$$\hat{Z}(x_0) = \sum_{i=1}^N w_i [Z(x_i) - \mu(x_0)] + \mu \quad (54)$$

Here,  $\hat{Z}(x_0)$  is the predicted parameter value at point  $x_0$ ,  $\mu$  is the constant mean value over the region of interest. The OK assumes the stationarity of the first moment of the prediction parameter, that is  $E\{Z(x_i)\} = E\{Z(x_0)\} = \mu = \mu(x_0)$ ,  $\mu$  is unknown. The  $w_i$  is measured from the experimental semivariogram.

$$\begin{aligned}\gamma(h) \\ = \frac{\sum_{i=1}^N [Z(x_i) - Z(x_i + h)]^2}{2M}\end{aligned}\quad (55)$$

Here,  $\gamma(h)$  represents the semi variance at lag interval  $h$ ,  $Z(x_i)$  is measured parameter value at point  $x_i$ ,  $Z(x_i + h)$  is measured parameter value at sampled location which is  $h$  lag distance apart from  $x_i$ ,  $M$  is the total number of pairs of the interpolating points that are  $h$  distance lag apart. Detail of semantic Kriging for spatial-temporal prediction can reference in studies (Bhattacharjee et al., 2019).

#### 2.10. Zonal statistic model

Zonal statistic methods summarize and aggregate the raster values intersecting a vector geometry. For instance, zonal statistics provide the mean precipitation or maximum elevation of an administrative unit. Additionally, functions are provided to query a raster at a point and get an interpolated value rather than the simple nearest pixel. The values within the zone were assumed normalized. The percentile was computed based on mean and standard deviation. The extreme value of zone use percentile method Q1 (Hyndman and Fan 1996) for query. The extreme drought values are determined when NDI equal or below to 10<sup>th</sup> percentiles or corresponds to  $z_{score} = -1.282$ . The mean ( $\mu$ ), standard deviation ( $\sigma$ ),  $z_{score}$  and percentile values of zone ( $x_p$ ) was computed follow equation (3, 4, 5, 6) below:

$$\mu = \frac{1}{N} \sum_i^N x_i \quad (56)$$

$$\sigma = \sqrt{\frac{1}{N} \sum_i^N (x_i - \mu)^2} \quad (57)$$

$$z_{score} = \frac{x - \mu}{\sigma} \quad (58)$$

$$x_p = z_{score} \times \sigma + \mu \quad (59)$$

### 2.11. Deep learning models

The deep learning model uses encoder-decoder model, combine with both LSTM, CNN.

The encoder-decoder architecture for recurrent neural networks is proving to be powerful on a host of sequence-to-sequence prediction problems in the field of natural language processing such as machine translation and caption generation. Attention is a mechanism that addresses a limitation of the encoder-decoder architecture on long sequences, and that in general speeds up the learning and lifts the skill of the model no sequence-to-sequence prediction problems. The model consists of 3 parts: encoder, intermediate (encoder) vector and decoder.

Encoder stacks of several recurrent units. Each step accepts a single element of the input sequence, collects information, and propagates to forward. The hidden states  $h_i$  are computed follow the equation:

$$h_t = f(W^{(hh)}h_{t-1} + W^{(hx)}x_t) \quad (60)$$

It is used to apply the appropriate weights to the previously hidden states  $h_{t-1}$  and input vector  $x_t$ .

Encoder vector is the final hidden state produced from the encoder part. This vector is used to encapsulate the information for all input elements to help the decoder make accurate predictions. It acts as the initial hidden state of the decoder part of the model.

Decoder stacks of several recurrent units where each predicts an output  $y_t$  at a time step t. Each recurrent unit accepts a hidden state from previous unit and produces an output as well as its own hidden state. Any hidden state  $h_i$  is computed follow equation:

$$h_t = f(W^{(hh)}h_{t-1}) \quad (61)$$

The output  $y_t$  at time step t is computed using the equation:

$$y_t = \text{softmax}(W^S h_t) \quad (62)$$

The output was calculated by using the hidden state at the current time step and respective weight  $W^S$ .

The power of this model lies in the fact that it can map sequences of different lengths to each other. The inputs and outputs are not correlated, and their lengths can differ. This opens a whole new range of problems which can be solved using such architecture.

The CNN-LSTM combines the advantages of CNN and LSTM. The  $m \times c$  variables at  $n$  location in the past  $t$  hours were taken as input data. The input dimension was the same as the CNN, which was  $t \times m \times n$ . The input convolution, pooling layers of CNN were used to extract the features of the input data. The obtained features were flattened into the 1-D array as the time sequence of sequential input data of LSTM. Finally, the prediction values at  $n$  locations in the next time  $t$  were obtained through the fully connection and output layers. The interpolated CNN-LSTM-AE model was presented in Fig 3.

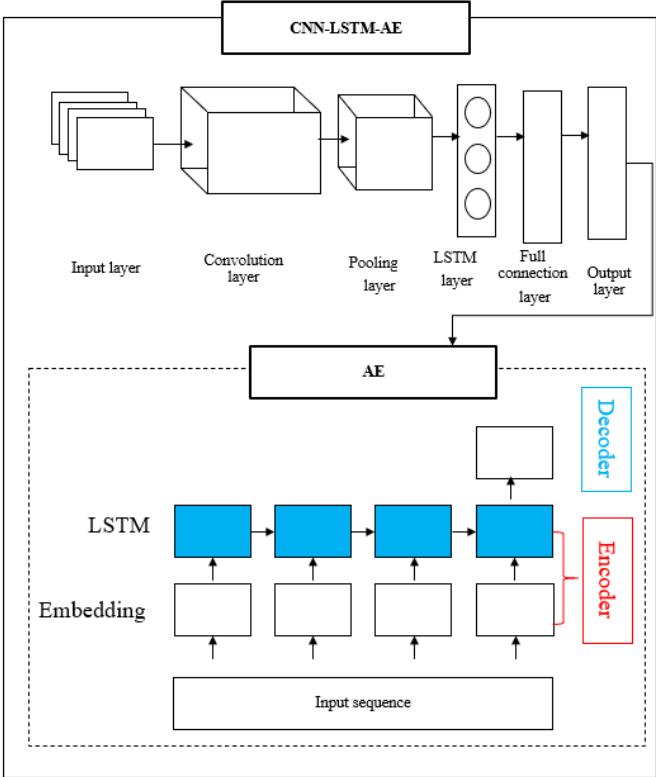


Fig 186. Interoperated CNN-LSTM-AE model.

## References

- Arnone, E., Cucchi, M., Gesso, S. D., Petitta, M., & Calmant, S. (2020). Droughts Prediction: a Methodology Based on Climate Seasonal Forecasts. *Water Resources Management*, 34(14), 4313-4328. doi:10.1007/s11269-020-02623-3
- Barzegar, R., Aalami, M. T., & Adamowski, J. (2020). Short-term water quality variable prediction using a hybrid CNN–LSTM deep learning model. *Stochastic Environmental Research and Risk Assessment*, 34(2), 415-433. doi:10.1007/s00477-020-01776-2
- Dikshit, A., Pradhan, B., & Alamri, A. M. (2021). Long lead time drought forecasting using lagged climate variables and a stacked long short-term memory model. *Science of The Total Environment*, 755, 142638. doi:<https://doi.org/10.1016/j.scitotenv.2020.142638>
- Ferreira, L. B., da Cunha, F. F., de Oliveira, R. A., & Fernandes Filho, E. I. (2019). Estimation of reference evapotranspiration in Brazil with limited meteorological data using ANN and SVM – A new approach. *Journal of Hydrology*, 572, 556-570. doi:10.1016/j.jhydrol.2019.03.028
- Fung, K. F., Huang, Y. F., Koo, C. H., & Soh, Y. (2020). Drought forecasting: A review of modelling approaches 200W 2017. *Journal of Water and Climate Change*, 11, 771-799.

- Gao, S., Wang, X., Miao, X., Su, C., & Li, Y. (2019). ASM1D-GAN: An Intelligent Fault Diagnosis Method Based on Assembled 1D Convolutional Neural Network and Generative Adversarial Networks. *Journal of Signal Processing Systems*, 91(10), 1237-1247. doi:10.1007/s11265-019-01463-8
- García-León, D., Standardi, G., & Staccione, A. (2021). An integrated approach for the estimation of agricultural drought costs. *Land Use Policy*, 100, 104923. doi:<https://doi.org/10.1016/j.landusepol.2020.104923>
- Goodfellow, I., Bengio, Y., Courville, A., & Bengio, Y. (2016). Deep learning (Vol. 1): MIT press Cambridge.
- Gupta, V., & Jain, M. K. (2021). Unravelling the teleconnections between ENSO and dry/wet conditions over India using nonlinear Granger causality. *Atmospheric Research*, 247, 105168. doi:<https://doi.org/10.1016/j.atmosres.2020.105168>
- Hao, Z., Singh, V. P., & Xia, Y. (2018). Seasonal drought prediction: advances, challenges, and future prospects. *Reviews of Geophysics*, 56(1), 108-141.
- Her, Y., Yoo, S.-H., Cho, J., Hwang, S., Jeong, J., & Seong, C. (2019). Uncertainty in hydrological analysis of climate change: multi-parameter vs. multi-GCM ensemble predictions. *Scientific Reports*, 9(1), 4974. doi:10.1038/s41598-019-41334-7
- Heryadi, Y. (2019, 2019//). Learning Hierarchical Weather Data Representation for Short-Term Weather Forecasting Using Autoencoder and Long Short-Term Memory Models. Paper presented at the Intelligent Information and Database Systems, Cham.
- Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, 313(5786), 504-507.
- Jehanzaib, M., Shah, S. A., Yoo, J., & Kim, T.-W. (2020). Investigating the impacts of climate change and human activities on hydrological drought using non-stationary approaches. *Journal of Hydrology*, 588, 125052. doi:<https://doi.org/10.1016/j.jhydrol.2020.125052>
- Kim, T.-Y., & Cho, S.-B. (2019). Predicting residential energy consumption using CNN-LSTM neural networks. *Energy*, 182, 72-81. doi:10.1016/j.energy.2019.05.230
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. doi:10.1038/nature14539
- Lee, T., Shin, J.-Y., Kim, J.-S., & Singh, V. P. (2020). Stochastic simulation on reproducing long-term memory of hydroclimatological variables using deep learning model. *Journal of Hydrology*, 582, 124540. doi:10.1016/j.jhydrol.2019.124540
- Liu, J. N., Hu, Y., He, Y., Chan, P. W., & Lai, L. (2015). Deep neural network modeling for big data weather forecasting. In *Information Granularity, Big Data, and Computational Intelligence* (pp. 389-408): Springer.
- Liu, Y., & Chen, J. (2021). Future global socioeconomic risk to droughts based on estimates of hazard, exposure, and vulnerability in a changing climate. *Science of The Total Environment*, 751, 142159. doi:<https://doi.org/10.1016/j.scitotenv.2020.142159>
- NOAA. (2020). Retrieved from [ncdc.noaa.gov/billions/](http://ncdc.noaa.gov/billions/)
- Sayeed, A., Choi, Y., Eslami, E., Lops, Y., Roy, A., & Jung, J. (2020). Using a deep convolutional neural network to predict 2017 ozone concentrations, 24 hours in advance. *Neural Networks*, 121, 396-408. doi:10.1016/j.neunet.2019.09.033
- Shen, C. (2018). A Transdisciplinary Review of Deep Learning Research and Its Relevance for Water Resources Scientists. *Water Resources Research*, 54(11), 8558-8593. doi:10.1029/2018WR022643

- Stahl, K., Kohn, I., Blauthut, V., Urquijo, J., De Stefano, L., Acacio, V., . . . Tallaksen, L. (2015). Impacts of European drought events: insights from an international database of text-based reports. *Natural Hazards & Earth System Sciences Discussions*, 3(9).
- Strazzo, S., Collins, D. C., Schepen, A., Wang, Q., Becker, E., & Jia, L. (2019). Application of a hybrid statistical-dynamical system to seasonal prediction of North American temperature and precipitation. *Monthly Weather Review*, 147(2), 607-625.
- Wilhite, D. A. (2000). Drought as a natural hazard: concepts and definitions.
- Wu, Z., Hou, B., & Jiao, L. (2020). Multiscale CNN With Autoencoder Regularization Joint Contextual Attention Network for SAR Image Classification. *IEEE Transactions on Geoscience and Remote Sensing*, 1-14. doi:10.1109/TGRS.2020.3004911
- Xu, L., Chen, N., Zhang, X., & Chen, Z. (2018). An evaluation of statistical, NMME and hybrid models for drought prediction in China. *Journal of Hydrology*, 566, 235-249. doi:<https://doi.org/10.1016/j.jhydrol.2018.09.020>
- Yao, N., Li, L., Feng, P., Feng, H., Li Liu, D., Liu, Y., . . . Li, Y. (2020). Projections of drought characteristics in China based on a standardized precipitation and evapotranspiration index and multiple GCMs. *Science of The Total Environment*, 704, 135245. doi:<https://doi.org/10.1016/j.scitotenv.2019.135245>
- Zhao, C., Brissette, F., Chen, J., & Martel, J.-L. (2020). Frequency change of future extreme summer meteorological and hydrological droughts over North America. *Journal of Hydrology*, 584, 124316. doi:<https://doi.org/10.1016/j.jhydrol.2019.124316>

### WEEK 3

#### Weekly report detail

This report reviews the recently published drought prediction studies using machine learning and extracted the idea how to evaluate forecast model.

#### 1. Review drought predicted studies.

We chose 6 papers from Advances in Water Resources, Journal of Hydrology, Journal of Water and Climate Change, Atmosphere, Sustainable Water Resources Management, Science of The Total Environments for reviewing. We extracted 5 things from each study: (1) How they simulate the drought phenomenon; (2) How they determine extreme drought; (3) Which models they use; (4) How they process data; (5) How they evaluate the model. The summary of results is shown in **Table 1** below.

*Table 24. Summary of drought prediction in various journals.*

No.	Articles	Objective	Methods
1	(Dikshit et al., 2020)	Forecast temporal hydrological drought	<ul style="list-style-type: none"> <li>①. Simulate drought: SPEI.</li> <li>②. Extreme drought: <math>\text{SPEI} \leq -2.0</math></li> <li>③. Forecast model: ANN, SVR model.</li> <li>④. Train and test model: train (1901-2010), test (2011-2018)</li> <li>⑤. Evaluate model: MAE, R<sup>2</sup> to compare ANN and SVR. ANN is better because R<sup>2</sup> is higher than SVR ANN is better because R<sup>2</sup> is higher than the SVR</li> </ul>
2	(Dikshit et al., 2021)	Forecast lead times of drought from 1 month to 12 months	<ul style="list-style-type: none"> <li>①. Simulate drought: SPEI.</li> <li>②. Extreme drought: <math>\text{SPEI} \leq -2</math></li> <li>③. Forecast model: LSTM.</li> <li>④. Train and test model: monthly SPEI train (1981-2000), test (2001-2018).</li> <li>⑤. Valuate model: coefficient of determination (<math>R^2</math>) and RMSE. Compare the mean gridded observed and forecasted SPEI follow TS (Jolliffe &amp; Stephenson, 2012):  <math display="block">TS = \frac{\text{hits}}{\text{hits} + \text{misses} + \text{falsealarms}}</math> </li> </ul>
3	(N. Khan et al., 2020)	Forecast the catalog drought (moderate, severe, extreme)	<ul style="list-style-type: none"> <li>①. Simulate drought: SPEI.</li> <li>②. Extreme drought: <math>\text{SPEI} \leq -1.5</math></li> <li>③. Forecast drought using SVM, ANN, KNN.</li> </ul>

No.	Articles	Objective	Methods
			<p>④. Train and test model: train (1948-1995), test (1996-2016).</p> <p>⑤. Evaluate model: compare <math>R^2</math>, NRMSE, PBIAS, md.</p>
4	(M. M. H. Khan et al., 2020)	Hybrid method for drought prediction	<p>①. Simulate drought: SPI.</p> <p>②. Extreme drought: <math>SPI \leq -2.0</math></p> <p>③. Forecast drought using ARIMA, ANN, WTF, W-ANN, ANN-ARIMA</p> <p>④. Train and test model: train (1948-1995), test (1996-2016).</p> <p>⑤. Evaluate model: <math>R^2</math>, NRMSE, PBIAS, md.</p>
5	(Belayneh et al., 2016)	Predict short term meteorological drought	<p>①. Simulate drought: SPI.</p> <p>②. Extreme drought: <math>SPI \leq -2.0</math></p> <p>③. Forecast drought using artificial neural networks (ANNs) and support vector regression (SVR), wavelet analysis (WA), coupled models (WA-ANN and WA-SVR).</p> <p>④. Train and test model: 80% train, 20% test (1970-2005)</p> <p>⑤. Evaluate model: RMSE, MAE, <math>R^2</math>.</p>
6	(Fung et al., 2019)	Predict lead one month drought	<p>①. Simulate drought: SPEI.</p> <p>②. Extreme drought: <math>SPEI \leq -2.0</math></p> <p>③. Forecast drought using support vector regression (SVR), Boosting-support vector regression (BS-SVR), Fuzzy-support vector regression (F-SVR).</p> <p>④. Train and test model: train (1976-2011), test (2012-2015)</p> <p>⑤. Evaluate model: MBE, RMSE, MAE, <math>R^2</math>.</p>

## **2. The extracted idea.**

We found that drought is often simulated by drought index and classified based on drought index categories. Machine learning studies usually split data into training and testing part. Training is for “teaching” model and “testing” for evaluating model with various metrics. To show the improvement, they compared several models together.

## **3. Clarify our study.**

- The objective: propose the framework to improve extreme drought prediction based on incorporated machine learning models. We use historical drought to predict the lead drought (1 month, 3 months, 6months, and 12 months).
- The natural drought Index is used to simulate the drought.
- Extreme drought is defined as NDI  $\leq -2.0$  with time series and 10<sup>th</sup> percentile for spatial distribution.
- 4 deep learning models (LSTM, encoder-decoder LSTM univariate, encoder-decoder LSTM multivariate, encoder-decoder CNN-LSTM, encoder-decoder Conv-LSTM) compare with the statistical model (ARIMA) to show the improvement of prediction.
- RMSE, MAE were used as the metrics.
- Expected results that will find better model (lower RMSE, MAE) for drought prediction.

## **4. Discussion**

(1) Beside review the articles, we also take the time to read the book for drought early warning and forecasting (Funk & Shukla, 2020). Chapter 10 of this book presents a practice-evaluating forecast skill. The skill score measures the performance of the forecasts during the past events. It assumes the current state of a variable will simply persist into the future over the target period. The target forecast will be like long-term average condition. The forecast is made based on pure change, with zero statistical or dynamic basis. For instance, in case, we use the historical extreme drought (2001-2016) as the “future” in the forecast

model. Then the short-term drought forecast skill score is evaluated based on the performance of it. Because the future condition is assumed “not change”.

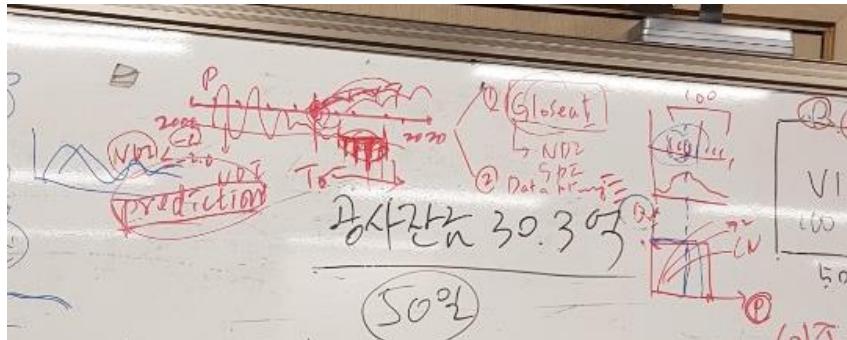
(2) The other viewpoint for estimate hydrology should consider in the interaction of dynamic landscape (Stephens et al., 2020). General hydrology is impacted and effect of human activity, natural environment. The predicted model should be analyzed with difference temporal, spatial scale, in the future landscape.

## References

- Belayneh, A., Adamowski, J., & Khalil, B. (2016). Short-term SPI drought forecasting in the Awash River Basin in Ethiopia using wavelet transforms and machine learning methods. *Sustainable Water Resources Management*, 2(1), 87-101.
- Dikshit, A., Pradhan, B., & Alamri, A. M. (2020). Temporal Hydrological Drought Index Forecasting for New South Wales, Australia Using Machine Learning Approaches. *Atmosphere*, 11(6), 585.
- Dikshit, A., Pradhan, B., & Alamri, A. M. (2021). Long lead time drought forecasting using lagged climate variables and a stacked long short-term memory model. *Science of The Total Environment*, 755, 142638. doi:<https://doi.org/10.1016/j.scitotenv.2020.142638>
- Fung, K. F., Huang, Y. F., Koo, C. H., & Mirzaei, M. (2019). Improved SVR machine learning models for agricultural drought prediction at downstream of Langat River Basin, Malaysia. *Journal of Water and Climate Change*, 11(4), 1383-1398. doi:10.2166/wcc.2019.295
- Funk, C., & Shukla, S. (2020). *Drought Early Warning and Forecasting: Theory and Practice*: Elsevier.
- Jolliffe, I. T., & Stephenson, D. B. (2012). *Forecast verification: a practitioner's guide in atmospheric science*: John Wiley & Sons.
- Khan, M. M. H., Muhammad, N. S., & El-Shafie, A. (2020). Wavelet based hybrid ANN-ARIMA models for meteorological drought forecasting. *Journal of Hydrology*, 590, 125380.
- Khan, N., Sachindra, D. A., Shahid, S., Ahmed, K., Shiru, M. S., & Nawaz, N. (2020). Prediction of droughts over Pakistan using machine learning algorithms. *Advances in Water Resources*, 139, 103562. doi:<https://doi.org/10.1016/j.advwatres.2020.103562>

Stephens, C., Lall, U., Johnson, F., & Marshall, L. (2020). Landscape changes and their hydrologic effects: Interactions and feedbacks across scales. *Earth-Science Reviews*, 103466.

Comments from professor:



- 1) Prepare PhD dissertation in the next sysmester
- 2) Your prosal up to now is not enough
- 3) It is not clear for the objective. Will you compare methods?
- 4) When you got results from different algorithm what would you say? Comprare different models for extreme drought is not enough.
- 5) Improve, extreme, prediction. How do you prediction?
- 6) In your description, I can not find any prediction. Our student use the gloasea5 , So, Kwan, Jaemin work in the dynamic model. ( weather forecasting model)
- 7) How to define the extreme drought after joint probability and distribution.
- 8) After copula, how do you say extreme will change or not? How do you difine extreme. ( theo minh thi phu thuoc vao return period)
- 9) You have no solution for prediction. Even you use crosify or data mining. You must make concept for prediction.
- 10) How extreme drought change from now up to 3 months. After movement one months, another forecastment.
- 11) Read Jaemin so or kwan at first. Those are physical based model. It is not nessacary all of them (3- 5 models). Chose the best.
- 12) Think about how do prediction
- 13) First understand simulation and prediction. All your describsion is no prediction.

#### Reply

- 1) Mục tiêu của nghiên cứu này là cải thiện dự báo hạn thông qua việc đề suất một mô hình máy học tổng hợp. Phát triển các ưu điểm của việc tổng hợp nhiều mô hình máy học trong việc dự báo hạn
- 2) Mô tả lại việc dự báo hạn theo chương 6 luận văn của A wan

## WEEK 4

### Weekly report detail

#### 5. Understand drought prediction.

Drought prediction based on various methods such as dynamic, statistical, hybrid and machine learning approach. The dynamic approach uses a physically based model such as land surface models, or climate model for prediction. The projected of precipitation, runoff from global climate model (ex, GloSea5) is used as the input data for physical based to forecast future runoff. This runoff is used to forecast drought (Bae et al., 2013; So et al., 2020; Son et al., 2015; Son & Bae, 2015). Fig 1 demonstrates the framework for real time drought prediction using GloSea5 data.

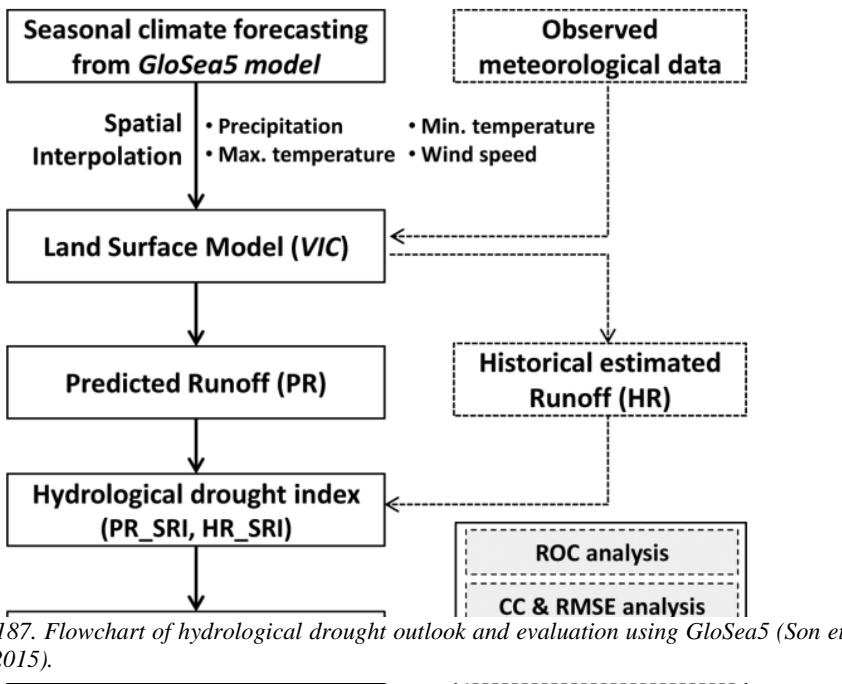


Fig 187. Flowchart of hydrological drought outlook and evaluation using GloSea5 (Son et al., 2015).

This study used historical runoff (HR) to forecast predicted runoff (PR) (Fig 2).

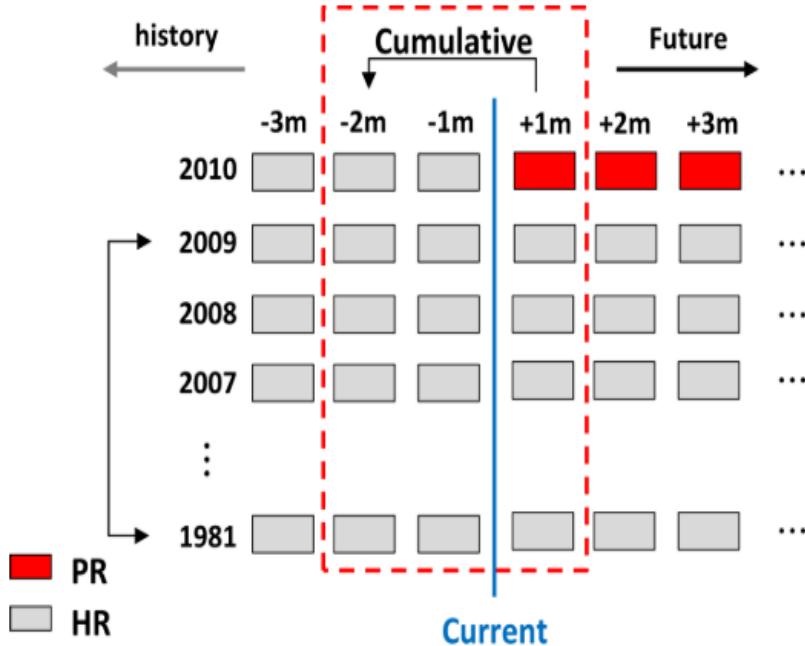


Fig 188. Drought outlook processing (Son et al., 2015).

Instead of Standardized Runoff Index (SRI), So et al. (2020) utilized Modified Surface Water Supply Index (MSWSI) for drought forecasts in South Korea. This study developed a hydrological drought forecasting linked to the land surface model. The data from GloSea5 was used to be input data from the VIC model to generate the Runoff forecasts, the anomaly ensemble forecast and hindcast data for each month was computed. The hybrid combines physical model and statistical model had been used to predict hydrological drought. General flow chart of drought forecast was presented in Fig 3.

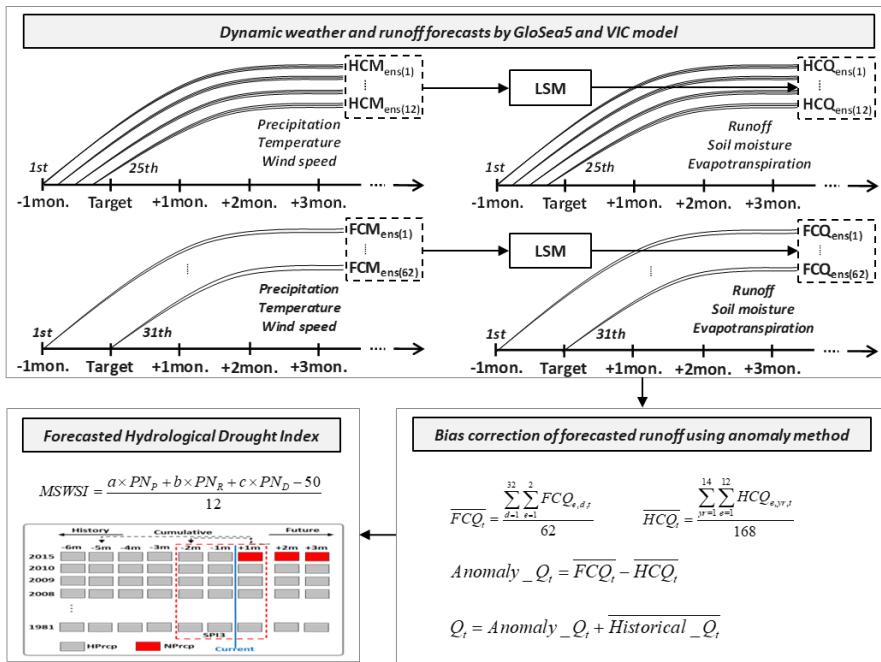


Fig 189. Outline of the hydrological drought forecasting framework. (a) procedure for the dynamic weather and runoff forecasts by GloSea5 and VIC model; (b) procedure for the bias correction of forecasted runoff using anomaly method; (c) procedure for the forecasted hydrological drought index.

Beside using the VIC model to simulate HR and Glosea5 data for hydrological drought prediction, The Bayesian method was utilized to improve drought prediction (Bae et al., 2017). Drought prediction based on the regression relationship between historical runoff (HR) and ensemble streamflow prediction runoff (ESP\_R) was applied. The dynamic prediction result produced through a coupled analysis of Global Seasonal Forecast System 5 (GS5) data and the LSM was applied in the likelihood function. The prediction framework was present in Fig 4 below.

● Starting point

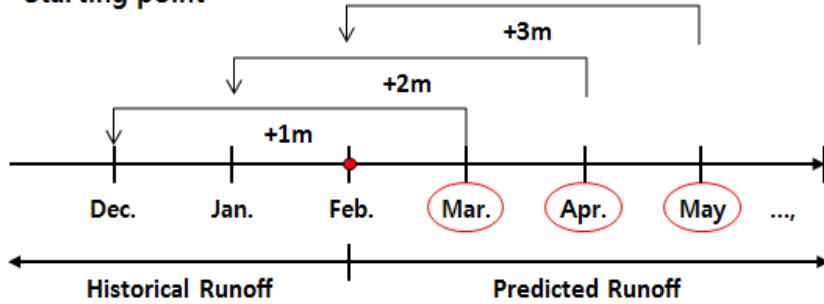


Fig 190. Schematic diagram of using the runoff data to estimate future SRI values for different lead times.

The meteorological drought prediction for East Asian monsoon was presented by Awan and Bae (2016). SPI was computed from the APHRODITE gridded rainfall. To improve drought prediction, Extended Reconstructed Sea Surface Temperature (ERSST) is also used as a predictor in the ANFIS model. Several models were compared to get the best configurations. The Drought prediction using historical data was presented in Table 1.

Table 25. Model with differential input and output for prediction.

Model	Inputs			Output
Model_A	SPI ( $t$ )			SPI ( $t+3$ )
Model_B	SPI ( $t-1$ )	SPI ( $t$ )		SPI ( $t+3$ )
Model_C	SPI ( $t-2$ )	SPI ( $t-1$ )	SPI ( $t$ )	SPI ( $t+3$ )
Model_D	SPI ( $t$ )	SSTA <sub>a</sub> ( $t$ )		SPI ( $t+3$ )
Model_E	SPI ( $t$ )	SSTA <sub>a</sub> ( $t$ )	SSTA <sub>b</sub> ( $t$ )	SPI ( $t+3$ )

This study uses various historical data from  $t$ ,  $t-1$ ,  $t-2$  and Sea Surface Temperature Anomaly (SSTA) to predict the drought.

## 6. Propose a drought prediction method.

The proposed framework for the extreme drought prediction was presented in Fig 5.

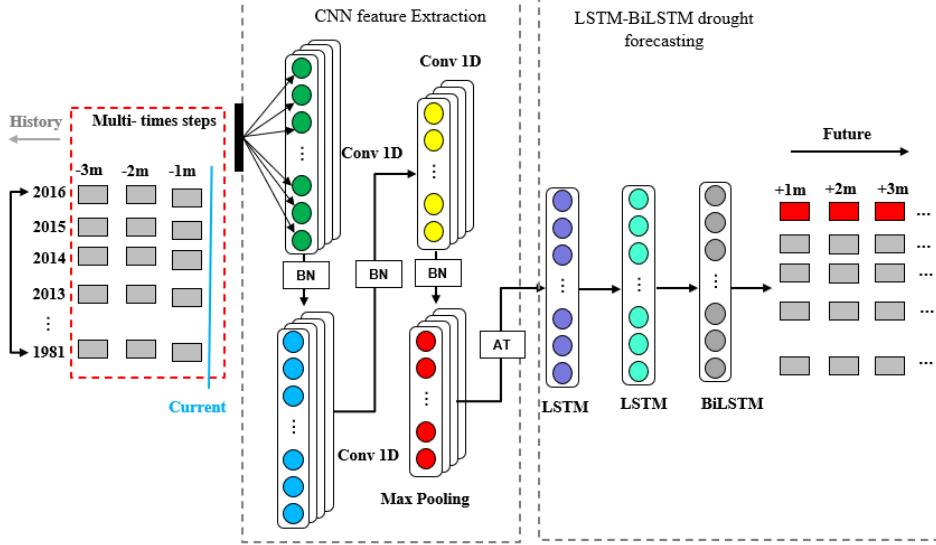


Fig 191. The algorithm of attention-based CNN-BiLSTM model.

The proposed model is composed of CNN, attention block and LSTM-BiLSTM, CNN extracts the features of the time series data, the attention block assigns weights to the features. LSTM-BiLSTM can forecast the load based on the extracted features. Each part of the proposed model is explained as follows.

1.CNN is designed with three one-dimensional convolutional layers, the number of convolutional kernels is set to 16, 32, 64, respectively. The size of convolutional kernel is set to 2, and the striding step is set to 1. Effective features can be extracted by striding the convolutional kernel on the features. After each convolution layer, the MaxPooling layer is added, the pooling window size is set to 2, and the striding step is set to 1. The MaxPooling layer can reduce the complexity of the features and avoid overfitting the model.

2.By assigning weights to features through the attention module, attention block can increase the impact of important time-series features on the model, suppress the interference of non-important features of the model, and effectively solve the problem that the model cannot judge the influence of the importance of different time series features.

3.The extracted features are taken as the input of two LSTM layers and one BiLSTM layer and obtain the NDI of the next month. Through multiple parameters fine tuning experiments, when the input historical data are chosen to the past 36 years, the trained model has the highest prediction accuracy, so the input is the monthly NDI from zones of the past 36 years.

## 7. Discussion.

These models above use climate model data, physical mode and statistical or artificial intelligent method for drought prediction. The prediction based on historical data that can be observed to predict the un-fore-seen drought events. Model were validated by using quantity methods. It compared to the collected drought information. We propose using only historical NDI data to predict drought. The extreme drought is defined as  $NDI \leq -2$ . The validation of model follows the extreme drought event in 2015-2016.

## References

- Awan, J. A., & Bae, D. H. (2016). Drought prediction over the East Asian monsoon region using the adaptive neuro-fuzzy inference system and the global sea surface temperature anomalies. *International Journal of Climatology*, 36(15), 4767-4777.
- Bae, D.-H., Ahn, J.-B., Kim, H.-K., Kim, H., Son, K.-H., Cho, S.-R., & Jung, U.-S. (2013). Development & evaluation of real-time ensemble drought prediction system. *Atmosphere*, 23(1), 113-121.
- Bae, D.-H., Son, K.-H., & So, J.-M. (2017). Utilization of the Bayesian method to improve hydrological drought prediction accuracy. *Water Resources Management*, 31(11), 3527-3541.
- So, J.-M., Lee, J.-H., & Bae, D.-H. (2020). Development of a Hydrological Drought Forecasting Model Using Weather Forecasting Data from GloSea5. *Water*, 12(10), 2785.
- Son, K.-H., Bae, D.-H., & Cheong, H.-S. (2015). Construction & evaluation of GloSea5-based hydrological drought outlook system. *Atmosphere*, 25(2), 271-281.
- Son, K. H., & Bae, D. H. (2015). Applicability assessment of hydrological drought outlook using ESP method. *Journal of Korea Water Resources Association*, 48(7), 581-593.

Professor comments:

- You should clarify to give the full proposal
- What is data you use
- How model is extracted

## **WEEK 5**

We don't have a meeting in this week. We should focus on how to write the methodology in this week. Understand models. How to processing data. Complete the proposal.

## WEEK 6

### Improve extreme drought prediction using incorporated deep learning models.

#### 8. Methodology

In this study we used the Natural drought index, Geostatistical and deep learning to improve drought prediction (Fig 1).

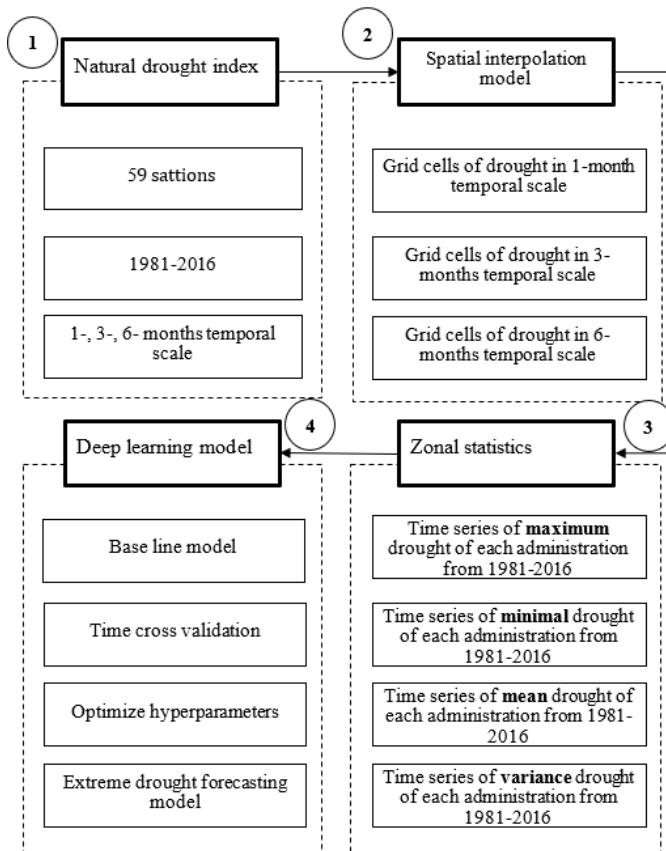


Fig 192. Flow chart of extreme drought prediction.

The detail of deep learning is presented in Fig 2. The first step is convolutional neural network model. The second is Long short-term memory. Final model is auto encoder decoder model.

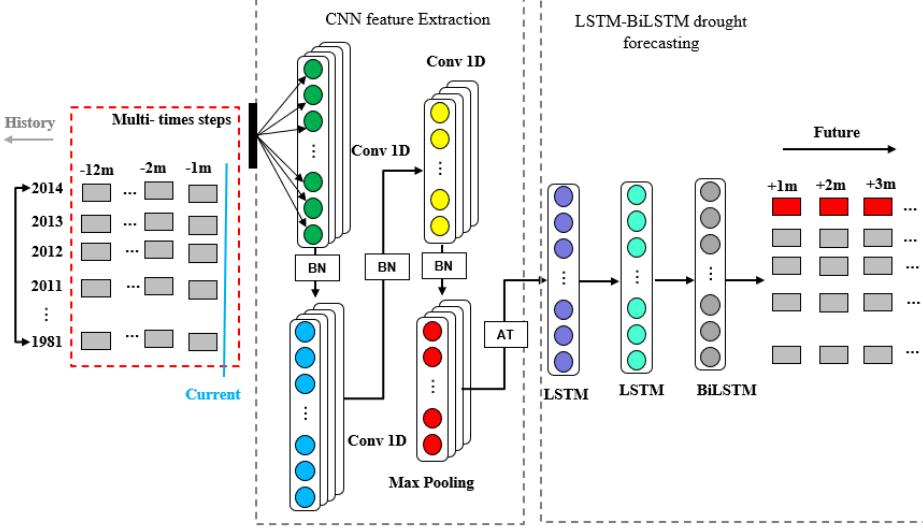


Fig 193. The algorithm of attention-based CNN-BiLSTM model.

General data from 1981-2014 is used to construct deep learning. The predict period is 2014-2015. We compare the drought characteristic including of number of drought events, drought duration, drought severity, drought interval between computed NDI and machine learning model. Then the spatial of them is examined.

For building a predictable model, we need to determine the input and output. The next predict values is impacted of method used for input. The fix window method means that input data used the last fix “3 months”, “6 months”, “12 months” to predict the next value. This method cannot model long-term dependencies. For instance, the extreme value in the last “15 months” does not impact to next prediction. The second idea that use entire sequence as set of counts that lead to the preserve order problem. It means that we use al value drought from 1981-2014 is different to data drought from 2014-1981. The other idea used a big, fixed window. It means we separate data, use one month input to one month prediction. The method has problem of parameter sharing.

To better prediction, a timeseries or model sequence should handle variable-length sequences, track long-term dependencies, maintain information about order, share parameters across the sequence. Because the seasonality of weather, we proposal use several schemes of annual period to determine optimal model. These schemes used various timescales to predict 1-, 3-, 6-, 12-, 24-months temporal scale of drought. This method is modified from the study of (Poornima & Pushpalatha, 2019).

## 9. Discussion

(1) We understood the prediction problem for sequence data. The predict method was setup.

However, the incorporation of multiple deep learning led to difficult to optimize hyperparameter. It required much more input parameter and the computational time of searching algorithm is increasing. Therefore, we proposed use the less complicated model of LSTM and Autoencoder. In Fig 3 presents one of typical LSTM-autoencoder model.

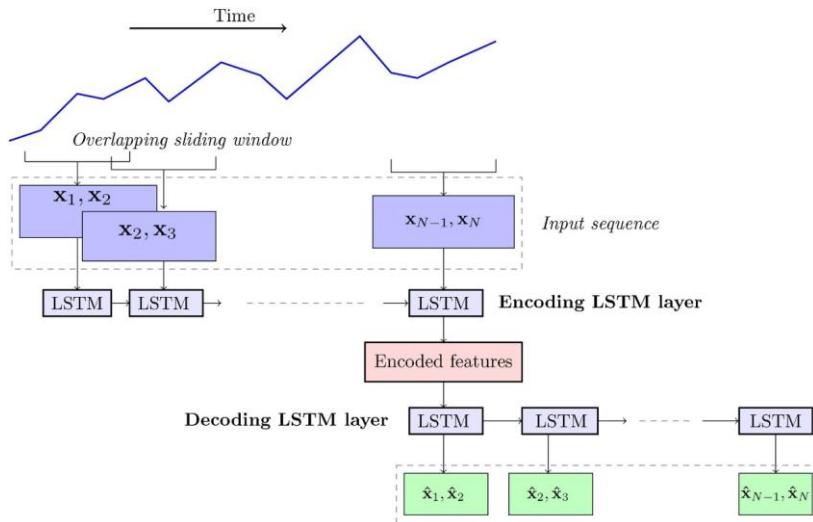


Fig 194. An illustration of the operation of the autoencoder LSTM network for the sliding window of size 2 (Nguyen et al., 2021). In our case the window size should be 3 months, 6 months, or 12 months to respect the seasonality of weather.

(2) To show the improvement of our model we propose compare with typical LSTM model that has been used popularly. The prediction in year 2015 and 2016 are used for comparison. Extreme drought characteristic included of number extreme drought events, drought duration, drought severity and drought interval will be compared.

## **References**

- Nguyen, H. D., Tran, K. P., Thomassey, S., & Hamad, M. (2021). Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management. *International Journal of Information Management*, 57, 102282. doi:<https://doi.org/10.1016/j.ijinfomgt.2020.102282>
- Poornima, S., & Pushpalatha, M. (2019). Drought prediction based on SPI and SPEI with varying timescales using LSTM recurrent neural network. *Soft Computing*, 23(18), 8399-8412. doi:10.1007/s00500-019-04120-1

## WEEK 7

### Improve extreme drought prediction using incorporated deep learning models.

We are analyzing 269 zones by Deep learning model. Total 432 months (1981-2016) NDI 3 data was split into 3 parts: training, validating, and predicting. 408 months (1981-2014) was used as training and testing. Ratio of training and testing is 0.8. Years 2015-2016 was chosen as the prediction. We built 4 schemes for drought prediction. The first scheme uses last 3 months NDI to predict the next values. The second, third, fourth schemes use last 6 months, 9 months, and 12 months to predict the next values.

Fig 1 presents the loss of training and testing for the first zone.

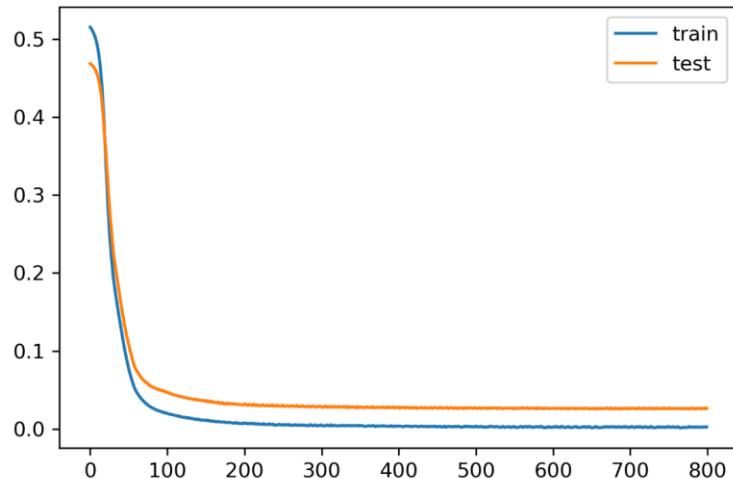


Fig 195. The loss of training and testing model for first zone.

The model shows the stable of learning rate with these selected parameters. It is lightly over-fitting cause of the test loss is higher than training. It is should be add dropout in the revised version. In addition, we suggest adding the early stopping layer to terminate training processing when the loss rate does not change.

Fig 2 presents the validation of model for one year. Noting that the RMSE, MAPE were computed for whole the testing period. It is not computed for only one year. We do not present the whole testing period causes of easy graphic visualization.

NDI of computation and Model simulation for one year at Andong106

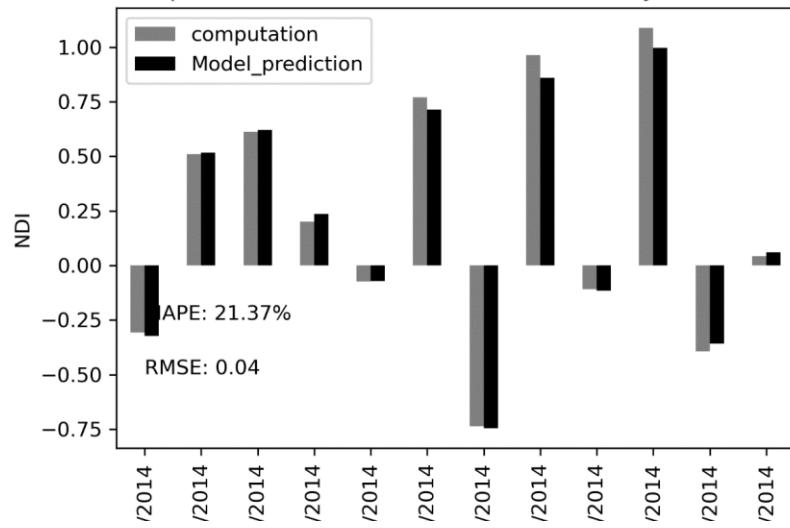


Fig 196. The Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) of model validation.

The model can achieve the trend and magnitude of drought and un-drought events.

Fig 3 presents the predicted drought for 2 years 2015-2016 at first zone.

NDI of computation and Model simulation for two year at Andong106

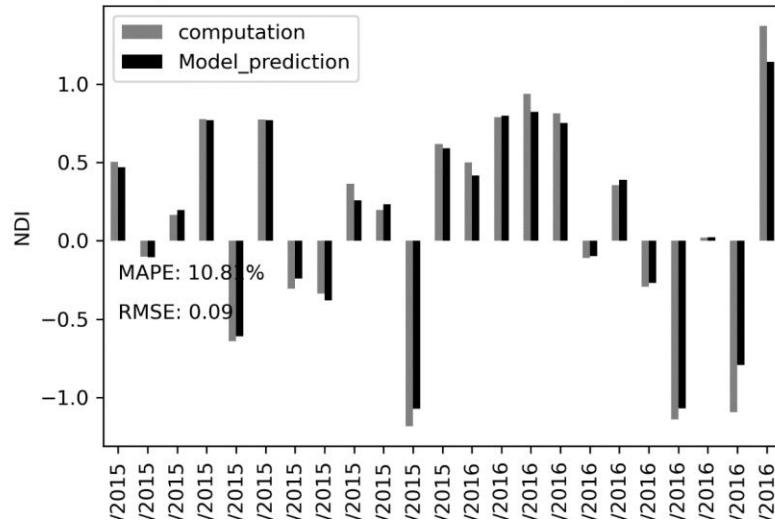


Fig 197. The Mean Absolute Percentage Error (MAPE) and Root Mean Square Error (RMSE) of model validation.

The MAPE and RMSE of predicted drought is lower than average of model validation. Because validation used several chunks of time series for assessing. Although this zone did not occur extreme drought, it still reflects the drought characteristic of “observation” and prediction. For instance, the dry occurred in October 2015 in both “observation” and prediction data.

The spatial distribution of predicted drought and observation will be mapped in the next steps after complete model simulation.

Professor comments:

-Draw 2 or 3 years validation

## WEEK 8

### Improve extreme drought prediction using incorporated deep learning models.

#### 10. Compare spatial distribution of predicted drought.

We created the spatial distribution of predicted drought from 269 zones for 24 months (2015-2016). Fig 1 presents the prediction drought map, observational drought map and the bias of prediction.

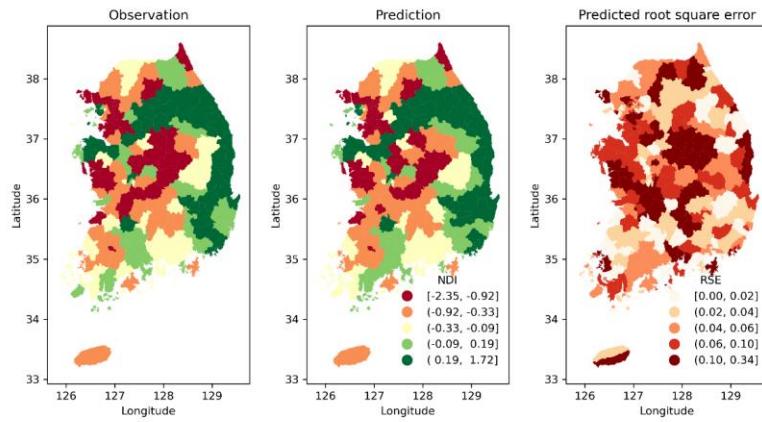


Fig 198. The observational, predicted and bias of spatial drought prediction.

Based on Fig 1, it shows the match of almost drought and extreme drought coverages. Drought occurred in the locations where is darker brown, while the wet occurred in locations where is yellow to green. The absolute bias is computed by root square of predicted and observational NDI values. The dark brown regions show the higher bias, while the light color shows more accurate predicted locations.

The mean absolute bias (MAE) and mean absolute percentage bias (MAPE) of spatial drought prediction for 24 months was presented in Fig 2.

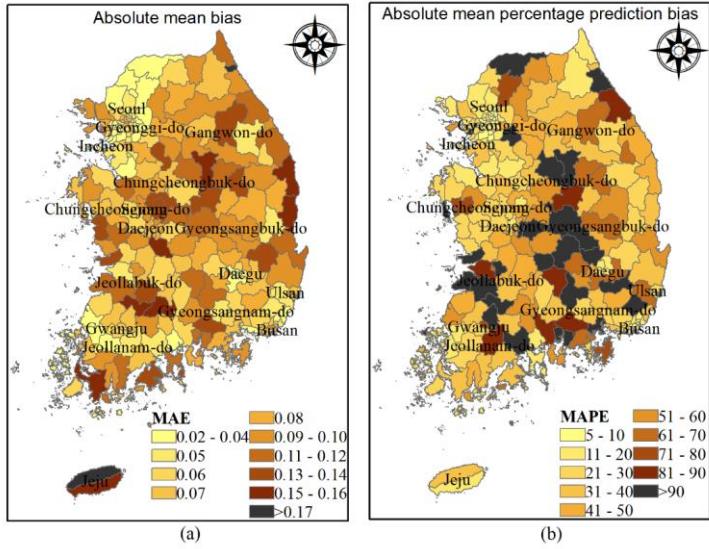


Fig 199. The mean absolute bias and mean absolute percentage bias of drought prediction.

Overall, the model is a good prediction at zones near Seoul, Gyeonggi-do, Busan, Jeollanam-do ( $MAPE < 10\%$ ). However, some zones have atrocious prediction near Gyongsangbuk-do, Jeollabuk-do, Daejeon, Incheon ( $MAPE > 90\%$ ).

Table 26. The mean absolute percentage bias of model prediction.

RANGE	TOTAL ZONES	%
[0 - 10.0]	17	6
(10.0 - 20.0]	70	26
[20.0 - 30.0]	53	20
[30.0 - 40.0]	33	12
[40.0 - 50.0]	22	8
[50.0 - 60.0]	19	7
[60.0 - 70.0]	12	4
[70.0 - 80.0]	5	2
[80.0 - 90.0]	7	3
(> 90.0]	31	12

In table 1 classifies the bias of spatial drought prediction. It shows that 64% zone has the bias between observation and prediction under 40%. The most zones have MAPE at 10-20%, and 31 zones corresponding 12 % of all zones have MAPE > 90%. Models should be adapted to improve the accuracy.

### 11. Increasing the speed of model training

We used the early stopping (ES) technique to decrease training time. The basic concept of the ES is a termination model when the gradient of model learning rate is unchanged, or lower than the pre-determined threshold. For instance, we set the number of epochs is 800, but the model is stopped at number epochs 145 when the loss is not much changed (Fig 3).

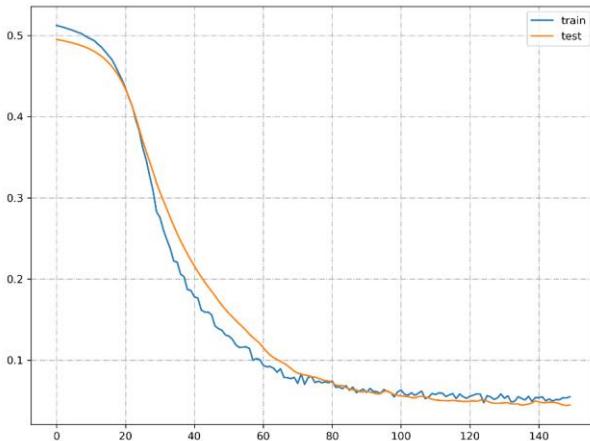


Fig 200. The training and testing model processing at the first zone (Andong 106).

The ES technique should be used carefully because it can because of local optimization. It means that the model does not reach the global optimum. Fig 4 presents the processing model dose not using ES technique.

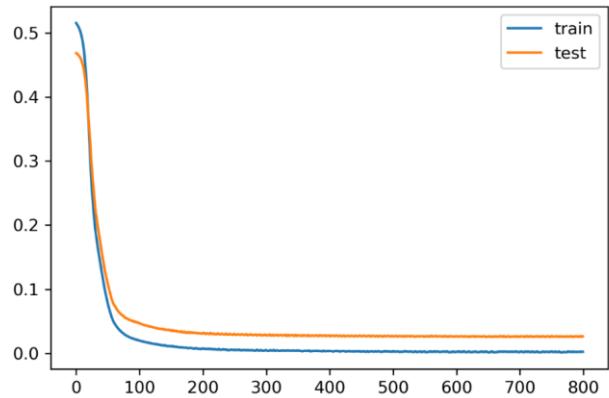


Fig 201. The training and testing model processing at the first zone (Andong 106) without ES.

Compare Fig 3 and Fig 4 we can see the “overfitting” was handled when the testing loss is not higher than training loss. It means that the validation of the model is obtained.

In the next step, we proposal finding the reason why model was poor prediction at some zones and how to improve the prediction at these zones.

## WEEK 9

**Improve extreme drought prediction using incorporated deep learning models.**

### 12. Finding the sources of error

We detached the bad predicted zones to find the possible sources of error. In the Fig 1 presents the typical zone 70 where is a high bias of observation and prediction.

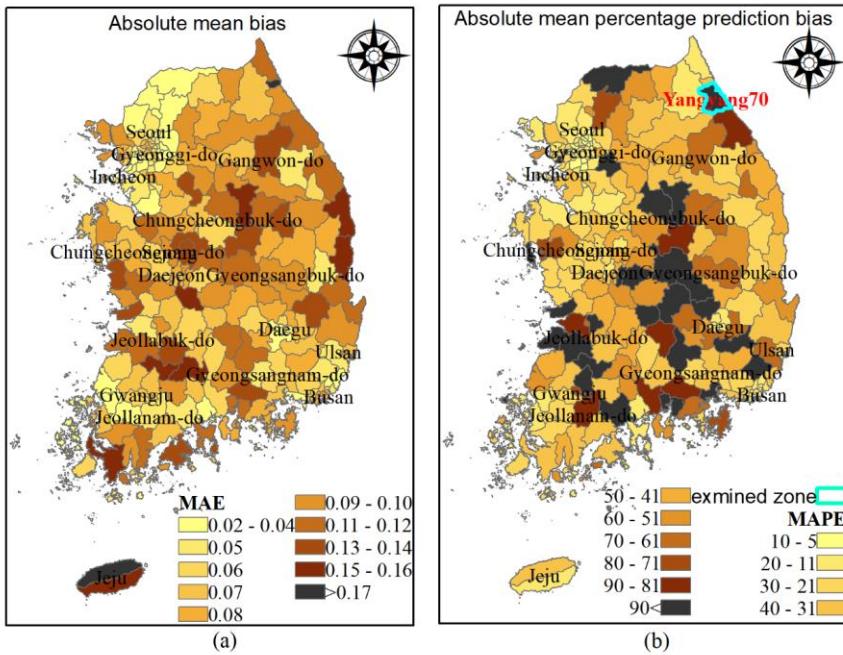


Fig 202. Examined zone is a high bias of prediction.

We look back the validation of model to check model parameter is suitable. Fig 2, Fig 3, Fig 4 below present the loss, validation, and prediction of zone 70.

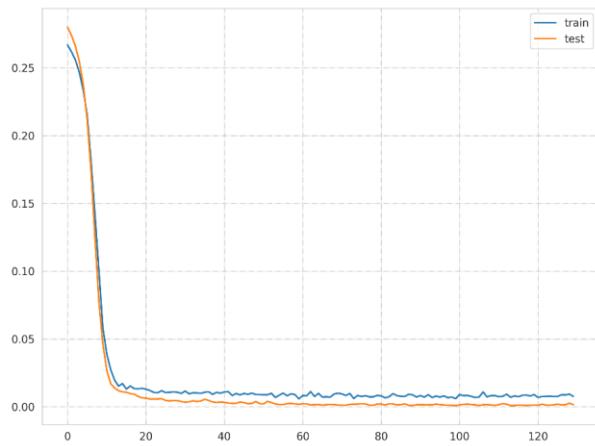


Fig 203. The loss of training and validation of model.

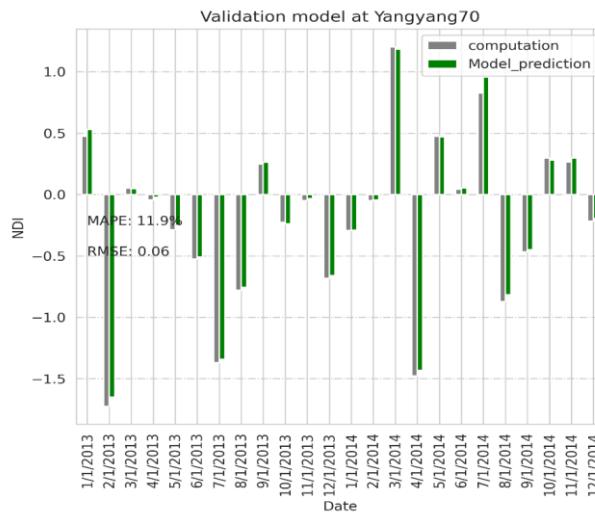


Fig 204. The validation of model from 2013-2014 where the original estimated data from 2008-2014.

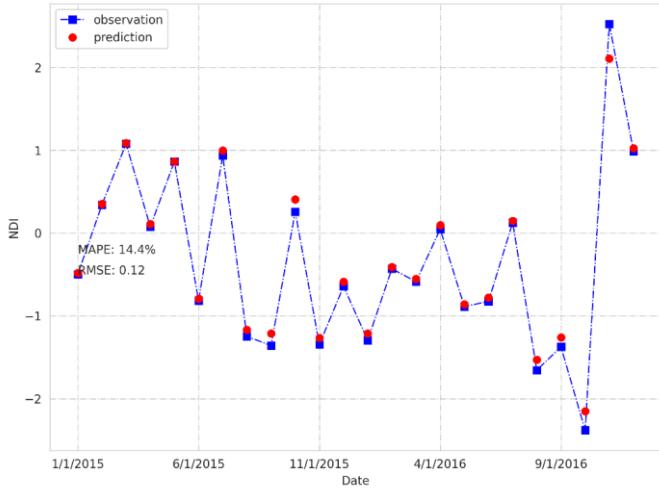


Fig 205. The model predicts drought at zone 70.

Results show that model is trained well, suitable parameters, and acceptable prediction at MAPE 14.4%. The reason of miss-matched name when we displayed error prediction on the map lead to error. The second reasons of error that model consist of wide range zones (269) that is needed for adaptive model parameters for some zones. Because it is hard to set one model fits for all.

### 13. Model adjustment

We propose using multiple models for this study. We had to check all the process and take the time to revise codes. Beside that we would like to find the solution for accelerating training model. Fig 5 presents the comparison error of model before and after adjustment. The model prediction has an improvement. The dark zones (high bias) are decreased. But they still exist at few zones.

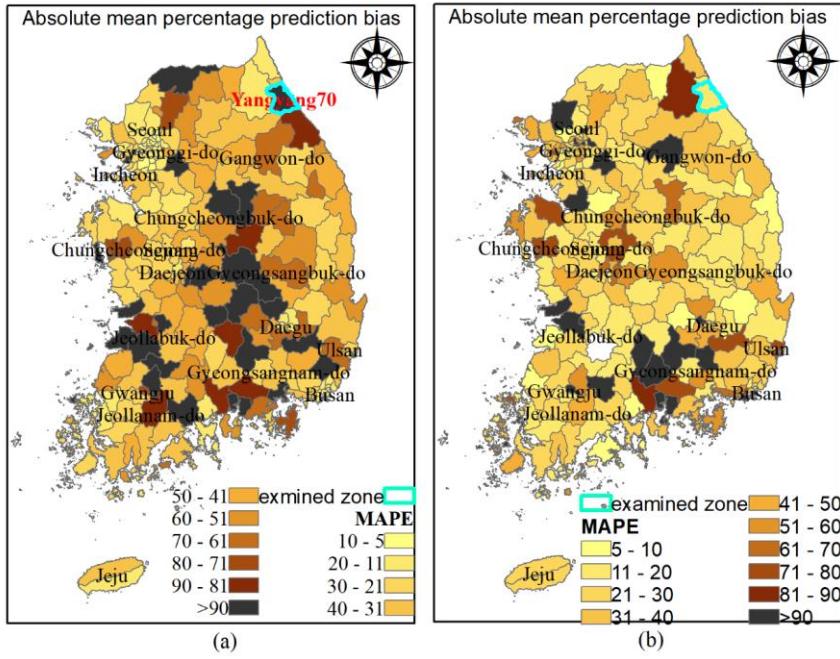


Fig 206. MAPE of model before (a) and after adjustment (b).

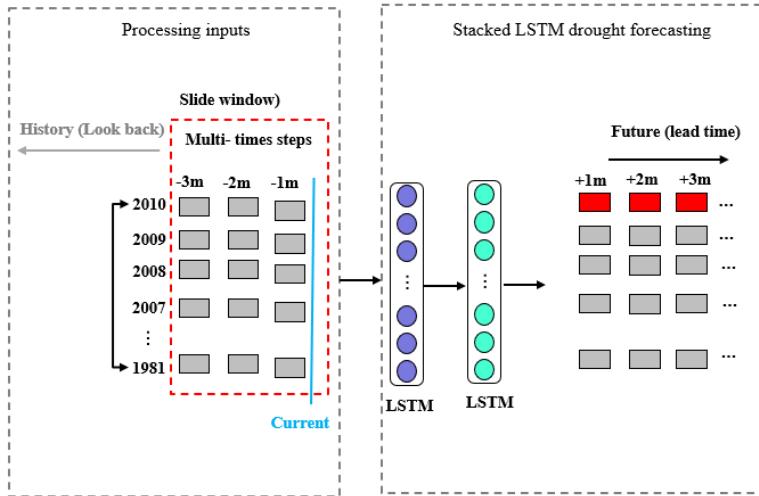
In the next step, we proposed checking the proposal algorithm and parameters, systematize and analyze results.

## WEEK 10

### Improve extreme drought prediction using incorporated deep learning models.

#### 14. Checking the proposal algorithms, and parameters.

The main algorithm is RNN which is adapted in various cases. It is preferring to supervising machine learning is used widen for prediction (Abayomi-Alli et al., 2019; Cai et al., 2019; Huang et al., 2020). We combined 4 “look back” schemes with 5 “lead time” schemes for prediction. The “look back” scheme defines number of previous steps are used for prediction. We used 12 months of the previous year to predict the next 3-, 6-, 12-, 24- months of drought. The leading time scheme define how long model will forecast after model was training/validation. It is various with prediction term above that it presents for total periods of drought prediction. For instance, we can set up the “look back” scheme using 12 months previously to get 3 months ahead as reformed of supervising machine. Then predict for 24 months later by slice the time window. Fig 1 present the algorithms of machine learning approach.



*Fig 207. The algorithm for drought prediction machine learning.*

Table 27. The summarize all of schemes.

No.	Look back	Window slide	Lead time
1	12	3	3
2	12	6	3
3	12	12	3

4	12	24	3
5	12	3	6
6	12	6	6
7	12	12	6
8	12	24	6
9	12	3	9
10	12	6	9
11	12	12	9
12	12	24	9
13	12	3	12
14	12	6	12
15	12	12	12
16	12	24	12
17	12	3	24
18	12	6	24
19	12	12	24
20	12	14	24

Figures (2 – 6) present the prediction of drought with lead time 3 months using “look back” 12 months and forward 3, 6, 12, 24 months setup.

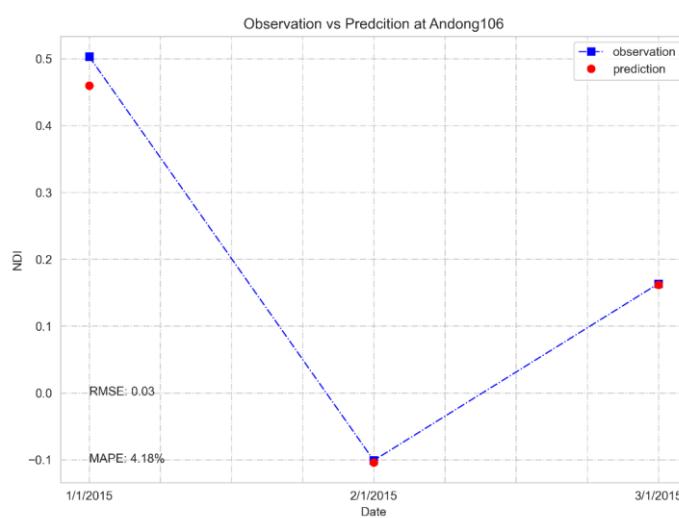


Fig 208. Predicted next three months based on “lookback” 12 months with slide window 3months.

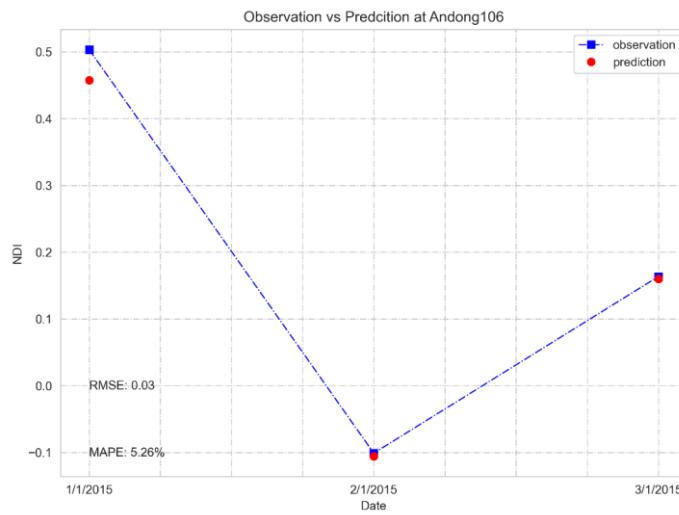
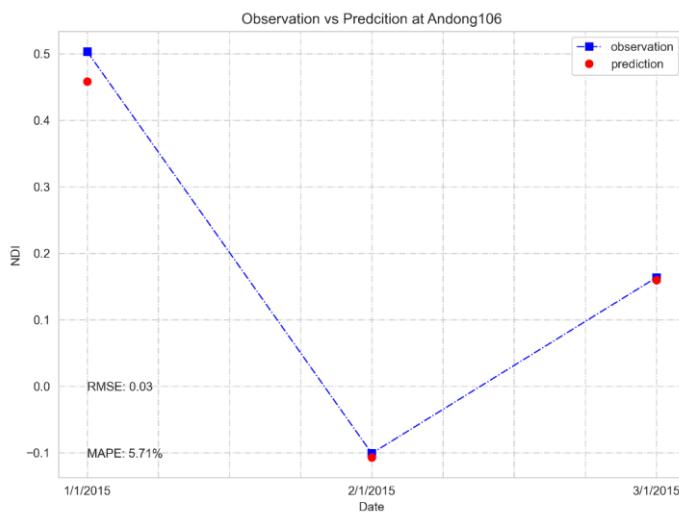
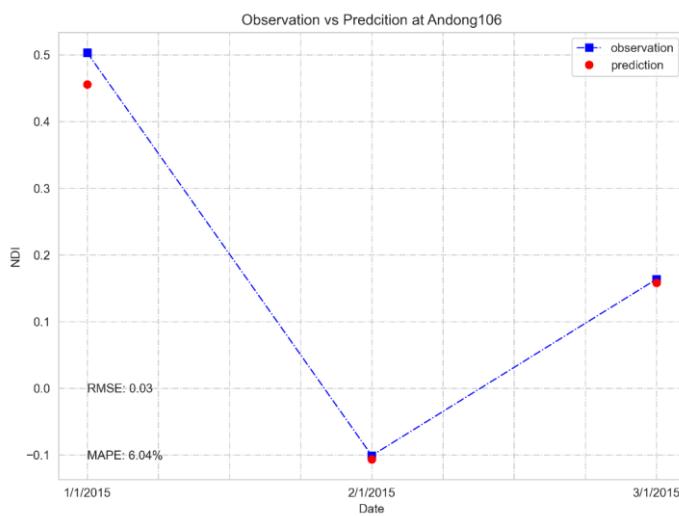


Fig 209. Predicted next three months based on “lookback” 12 months with slide window 6 months.



*Fig 210. Predicted next three months based on “lookback” 12 months with slide window 12 months.*



*Fig 211. Predicted next three months based on “lookback” 12 months with slide window 24 months.*

Logically, model is increasing the bias when we increase the window slice steps. Fig 1 has the smallest bias (MAPE = 4.18 %). While the side windows at 24 months, model has a greater bias (MAPE = 6.04%).

The accuracy of model depends on the lead time. At the same ‘lookback’ 12 months slide window 3 months but various lead times (3, 6, 12, 24 months) give various results. The longer lead times show the less accuracy results (Fig 6 to Fig 9).

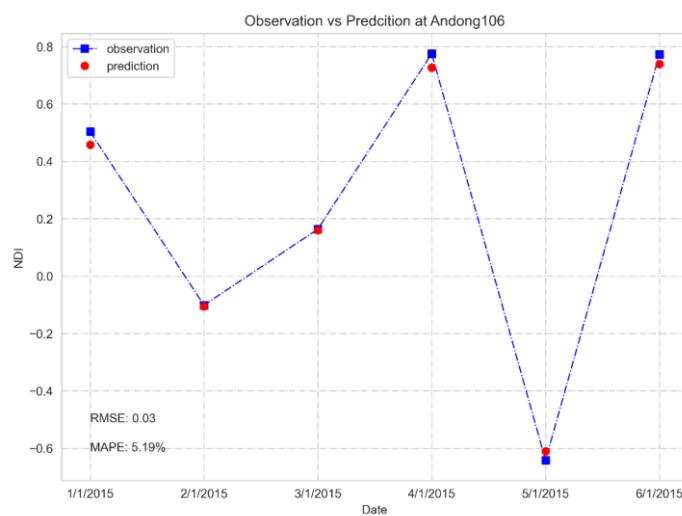


Fig 212. “lookback” 12, window slide 3 months, lead time 6 months prediction.

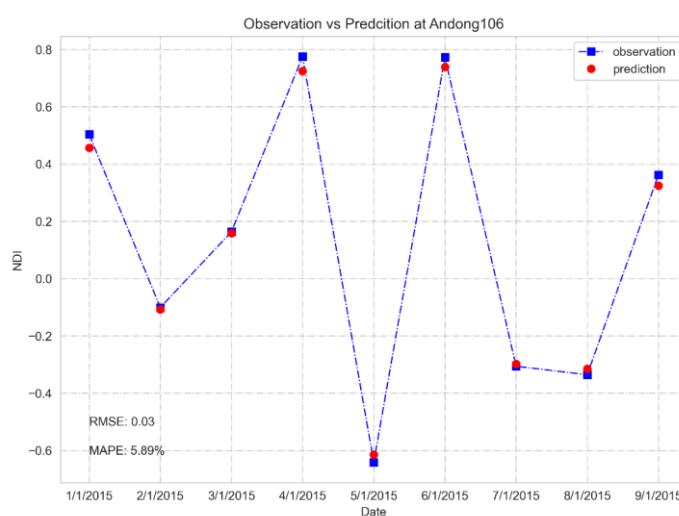


Fig 213. “lookback” 12, window slide 3 months, lead time 9 months prediction.

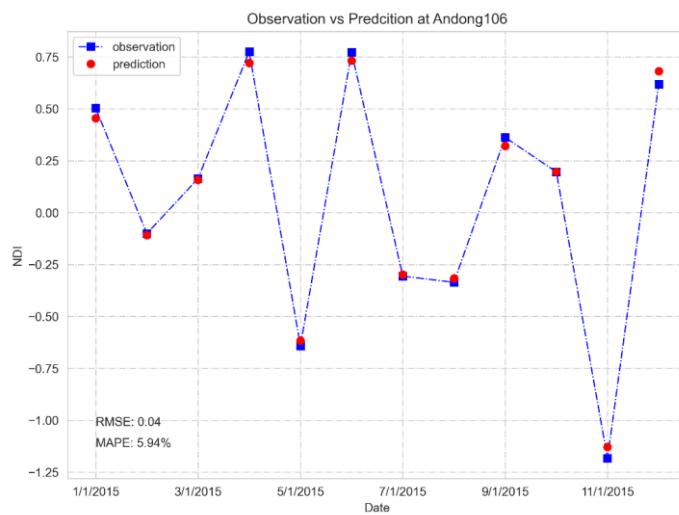


Fig 214. “lookback” 12, window slide 3 months, lead time 12 months prediction.

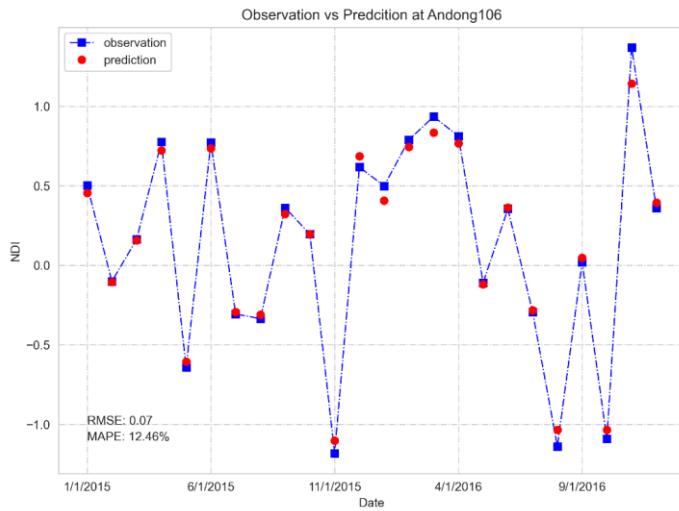


Fig 215. “lookback” 12, window slide 3 months, lead time 24 months prediction.

At the specific lead time scale, we analyze the best slide window with the same “lookback”. Results from Fig 10 to Fig 12 show widow slide 12 months is the best for 24 months drought prediction.

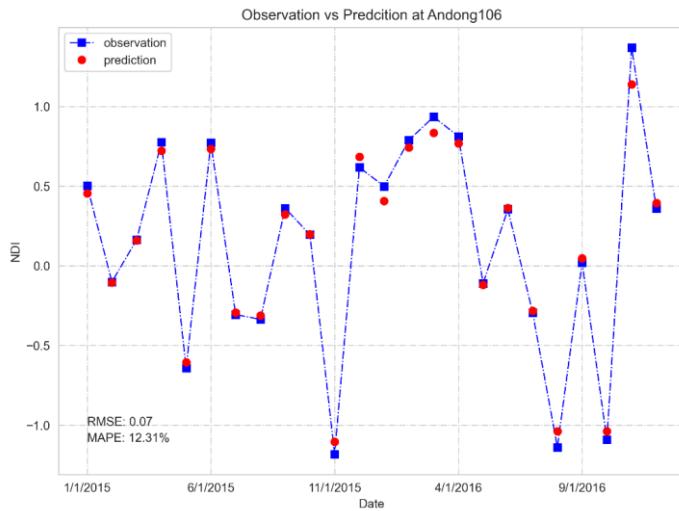


Fig 216. “lookback” 12, window slide 6 months, lead time 24 months prediction.

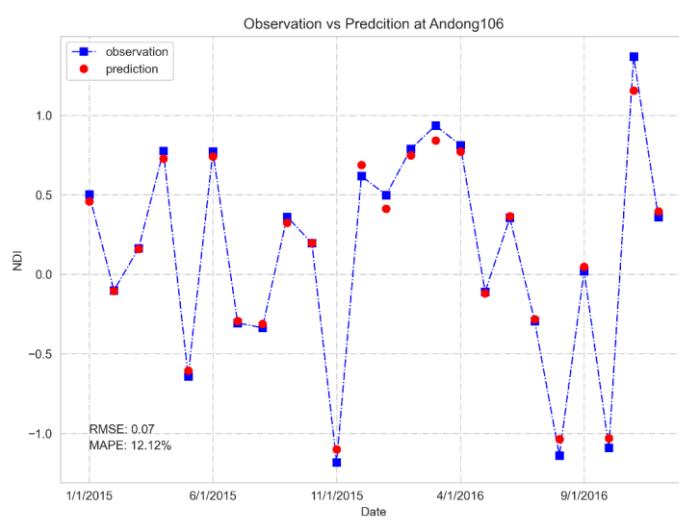


Fig 217. "lookback" 12, window slide 12 months, lead time 24 months prediction.

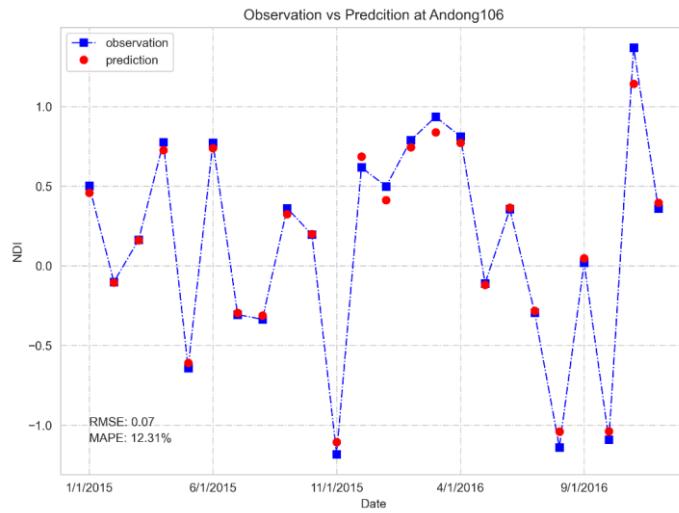


Fig 218. "lookback" 12, window slide 24 months, lead time 24 months prediction.

## **15. Discussion**

(13) The prediction algorithm was clarifying ‘lookback’, ‘widow slide’ and ‘lead time’ terms in our framework. Results shows that the lead time is increase make the bias is greater, and widows slide depended on the lead times. Because it determines how move the window to capture historical characteristic. It like to autofill function as the Microsoft excel works. If you choose (1,2,3) the next cell is 4, 5, 6. In that case is unchanged gradient, learning rate based on linearity. In the Machine learning, the activate function give the nonlinearity and then use loss compare the bias and optimizer to get the reliable reason for non-linearity cases. Some special case, model was added functions to better memorize the characteristic of history as forget gate, output gate in LSTM, encoder, decoder in the Attention based models.

(14) For passion of diving to machine learning, it could visualize the weights, bias of training of machine learning model to track the processing. Or we can customize the loss functions to get the suitable results, depended on the case by case. All results about were simulated with customized loss function Huber loss (Chen et al., 2017).

## **References**

- Abayomi-Alli, A., Odusami, M. O., Abayomi-Alli, O., Misra, S., & Ibeh, G. F. (2019). *Long short-term memory model for time series prediction and forecast of solar radiation and other weather parameters*. Paper presented at the 2019 19th International Conference on Computational Science and Its Applications (ICCSA).
- Cai, M., Pipattanasomporn, M., & Rahman, S. (2019). Day-ahead building-level load forecasts using deep learning vs. traditional time-series techniques. *Applied Energy*, 236, 1078-1088.
- Chen, C., Yan, C., Zhao, N., Guo, B., & Liu, G. (2017). A robust algorithm of support vector regression with a trimmed Huber loss function in the primal. *Soft Computing*, 21(18), 5235-5243. doi:10.1007/s00500-016-2229-4
- Huang, L., Zou, F., & Gan, Z. (2020). *Short Term Prediction of Colleges Building Itemized Energy Consumption Based on Long Short-term Memory Neural Network*. Paper presented at the 2020 IEEE 3rd International Conference of Safe Production and Informatization (IICSPI).

