**Dự án Danida**

**Nghiên cứu thuỷ tai do biến đổi khí hậu và xây dựng hệ thống thông tin nhiều bên tham gia nhằm giảm thiểu tính dễ bị tổn thương ở Bắc Trung Bộ Việt Nam (CPIS)**

Mã số . 11-P04-VIE

---

**Tên đề tài:**

# Dự án Nghiên cứu thuỷ tai do biến đổi khí hậu và xây dựng hệ thống thông tin nhiều bên tham gia nhằm giảm thiểu tính dễ bị tổn thương ở Bắc Trung Bộ Việt Nam

***Chủ nhiệm dự án: GS. TS. Phan Văn Tân***

**Báo cáo WP3:**

# BÁO CÁO KHOA HỌC VỀ KẾT QUẢ DỰ TÍNH KHÍ HẬU TƯƠNG LAI, TÍNH BẤT ĐỊNH VÀ PHƯƠNG PHÁP ĐÁNH GIÁ, XỬ LÝ

***Người thực hiện:*** ***Ngô Đức Thành***

# Nội dung 1.1:

# Báo cáo kĩ thuật chỉnh lý và định dạng số liệu nhiệt độ trung bình về dạng chuẩn các trạm Nghệ An - Hà Tĩnh – Quảng Bình

***Người thực hiện:***

## 1. Mở đầu

Hiện nay, các số liệu thu thập được từ mạng lưới quan trắc khá thô sơ và không đồng nhất. Số liệu quan trắc thu được còn ẩn chưa nhiều sai số làm cho khả năng phân tích các hiện tượng thời tiết cực đoan được tính toán dựa trên tập số liệu này vẫn còn yếu kém và không toàn diện nên thật khó đánh giá tần số các hiện tượng thời tiết cực đoan một cách chính xác. Do đó, các số liệu thu thập được cần chỉnh lý và định dạng lại theo một dạng chuẩn . Hơn nữa, với khả năng tính toán ngày càng cao của máy tính, chúng ta có thể xây dựng và phát triển chương trình định dạng và chuẩn hoá một cách tự động cũng như kiểm tra sai số quan trắc. Các phương pháp xác định sai số dựa trên phân bố không gian, thời gian cũng như tính đồng nhất giữa các biến. Sau đây chúng tôi sẽ mô tả phương pháp chỉnh lý và định dạng số liệu về dạng chuẩn cũng như phương pháp kiểm tra sai số.

## 2. Chỉnh lý và định dạng số liệu nhiệt độ trung bình về dạng chuẩn

Chương trình sau sẽ định dạng tập số liệu quan trắc thu thập được tự dạng ban đầu về dạng chuẩn, trong đó, số liệu sẽ được lưu trong cả giai đoạn trên tất cả các trạm và tất cả các biến, trong đó có biến nhiệt độ trung bình.

Từ cơ sở dữ liệu này, chương trình kiểm tra sai số sẽ tự động tìm các lỗi và nội suy giá trị thay thế nhằm đảm bảo tính hợp lý của giá trị về cả không gian, thời gian và phù hợp giữa các biến.

Chương trình được viết bằng ngôn ngữ fortran 90

Chương trình định dạng số liệu nhiệt độ trung bình về dạng chuẩn

```
Program Create_data
Integer, Parameter :: NVung=7, NSta=14, NYr1=1961, NYr2=2007, NDay=31, NYto=8, M=12
Integer, Parameter :: NtrMx = 200
Character *3, Parameter :: Var(NYto) = (/"T2m", "Tx", "Tm", "R", "BH", "Um", "U13", "Vx"/)
Character *11 :: StaName (NVung, NSta)

Character *3 :: YT(NYto)

Character (Len = 30) :: FInp, FOut, FOut1, Fmt
Character (Len = 16) :: TenVung(NVung)
Character (Len = 11) :: TenTram(NtrMx), VarName(NtrMx), Sample(Nmax)

Integer :: NTram
```

```
Real    :: X(Nmax, NtrMx), W1(Nmax)
Real*8  :: XX(Nmax, Mmax), X1(Nmax, Mmax), XX_R(Nmax,Mmax)
Integer :: YYDD(Nmax), YY(Nmax)
Real    :: Con

!-------------------------------------------------------------------------------
Real, Dimension (NVung, NSta, NYto, NYr1:NYr2, M, NDay) :: Dat        ! Data array

Integer (Kind=1), Dimension (NVung, NSta, NYr1:NYr2, M, NDay) :: &
                        H35, H37, HDry, HDry37, C15, C13, R25, R50

Integer, Dimension (NVung, NSta, NYr1:NYr2, M+1) :: NH35, NH37, NHDry, NHDry37, NC15, NC13, NR25, NR50, NWork
Real :: R_per(NVung*NSta, M), Per  ! Luu Phan vi Per cua mua
Real :: Rain_Dat(NVung*NSta, M, (NYr2-NYr1+1)*NDay
Real :: W((NYr2-NYr1+1)*NDay)  ! Mang lam viec

Character *11 DSTram(NVung*NSta)
Integer :: N_All_Sta
Dat(:,:,:,:,:,:) = -99.0

Open (4,file="SL_ALL.dat", form="Unformatted")

print*," Reading..."
Do iyt=1,NYto
  read(4) YT(iyt)
  Do iv=1,Nvung
    read(4) TenVung(iv)
    Do itr=1,NSta
      read(4) StaName(iv,itr)
      Do iyr= NYr1,NYr2
        read (4) iyear
        Do iday=1,Nday
          read(4) idate, (Dat(iv,itr,iyt,iyr,imon,iday),imon=1,M)
        Enddo
      Enddo
    Enddo
  Enddo
Enddo
Close(4)
! Tinh percentile cua Rain
Per = 0.90
Call Cal_Percentile(4, Per, "Percentile_R_Obs.txt")
! Tinh percentile cua Tx
Per = 0.90
Call Cal_Percentile(2, Per, "Percentile_Tx_Obs.txt")
! Tinh percentile cua Tm
Per = 0.10
Call Cal_Percentile(3, Per, "Percentile_Tm_Obs.txt")
```

## 3 Phương pháp kiểm tra và chất lượng số liệu quan trắc

Trong mọi trường hợp, số lượng (dung lượng mẫu) và chất lượng số liệu quan trắc, đặc biệt là những quan trắc tại trạm, có vai trò quyết định đối với kết quả nghiên cứu, tính toán và nhận định. Vì nhiều lí do khác nhau, nói chung các tập số liệu quan trắc đều tiềm ẩn các dạng sai số. Những giá trị quan trắc bất hợp lý nếu không được kiểm soát sẽ tác động đến những kết quả tính toán, phân tích và có thể dẫn đến những kết luận vô nghĩa. Bởi vậy, trước khi sử dụng các tập số liệu này cần thiết phải tiến

hành kiểm tra, đánh giá và xử lí những trường hợp nghi ngờ. Vì vậy trong nghiên cứu này sẽ thực hiện ba kiểm tra chất lượng sau:

- Kiểm tra khí hậu so sánh quan trắc với giá trị khí hậu. Kiểm tra này cũng thực hiện phần việc của kiểm tra vật lý do các giá trị ngưỡng được xác định dựa theo giá trị khí hậu riêng cho từng trạm.
- Kiểm tra phù hợp xác định tương thích về mặt vật lý giữa hai hay nhiều đại lượng.
- Kiểm tra không gian so sánh giá trị quan trắc với giá trị quan trắc từ các trạm xung quanh.

Kiểm tra khí hậu dựa trên đặc điểm khí hậu tại một khu vực xác định thông tin quan trắc có hợp lý hay không. Ví dụ nếu trong tháng bảy, ta nhận được một quan trắc nhiệt độ tối cao 20°C tại Hà Nội thì nhiều khả năng đây là một quan trắc sai bởi vào mùa hè, nhiệt độ tối cao thông thường tại Hà Nội vào khoảng 35°C với sai số 4°-5°C. Kiểm tra này có độ tin cậy cao với các biến số như nhiệt độ hay độ ẩm nhưng cần thận trọng khi sử dụng cho mưa bởi một quan trắc mưa lớn hơn rất nhiều giá trị khí hậu hoàn toàn có khả năng xuất hiện.

Thông thường, để xác định trung bình khí hậu và độ lệch chuẩn, ta có thể sử dụng các công thức thường dùng trong thống kê. Các công thức này có nhược điểm nếu xuất hiện một số hạng bất thường trong tập mẫu thống kê, giá trị trung bình và độ lệch chuẩn thống kê sẽ bị sai lệch hoàn toàn (ví dụ trong chuỗi nhiệt độ tháng bảy có số hạng 1000°C). Do đó, người ta thường dùng các đại lượng khác như Median và MAD để thay thế cho giá trị trung bình và độ lệch chuẩn khi xác định các đặc trưng thống kê. Trong dự án này, chúng tôi sử dụng phương pháp trung bình và độ lệch chuẩn hai trọng số của Lanzante (1996). Phương pháp này đã được một số tác giả sử dụng như Gleason (2002), Feng và nnk (2004). Trình bày dưới đây chủ yếu dựa theo Feng và CS (2004).

Theo phương pháp hai trọng số, các phần tử tập trung quanh tâm của phân bố sẽ có trọng số lớn hơn so với các phần tử bất thường nằm ngoài trung tâm khi tính toán giá trị trung bình và độ lệch chuẩn. Trọng số sẽ giảm dần đến 0 khi vượt quá một ngưỡng nào đó. Giả sử ta có một chuỗi các quan trắc $X_i$ gồm $n$ phần tử, có thể chứa các số hạng bất thường. Trọng số $u_i$ cho mỗi phần tử $X_i$ được tính như sau :

$$u_i = \frac{X_i - M}{c \times \mathrm{MAD}} \qquad (2.3.1)$$

với M, MAD là Median và MAD xác định từ chuỗi $X_i$, $c$ là hằng số cho biết ngưỡng loại bỏ các số hạng bất thường. Khi $u_i$ có giá trị tuyệt đối lớn hơn 1, nó sẽ được gán bằng 1, bảo đảm $u_i$ chỉ nằm trong khoảng [-1,1]. Điều này hoàn toàn được quyết định bởi giá trị của $c$ và $c$ được lấy bằng 7.5 theo Lanzante (1996). Giá trị trung bình và độ lệch chuẩn của chuỗi $X_i$ bây giờ được xác định như sau:

$$\overline{X}_{bi} = M + \frac{\sum_{i=1}^{n} (X_i - M)(1 - u_i^2)^2}{\sum_{i=1}^{n} (1 - u_i^2)^2} \qquad (2.3.2)$$

$$S_{bi} = \frac{\left[ n \sum_{i=1}^{n} (X_i - M)^2 (1 - u_i^2)^4 \right]^{0.5}}{\left| \sum_{i=1}^{n} (1 - u_i^2)(1 - 5u_i^2) \right|} \qquad (2.3.3)$$

nghĩa là $(1-u_i^2)$ đóng vai trò trọng số thay vì $u_i$. Đây là một đánh giá tốt của giá trị trung bình và độ lệch chuẩn, không chịu tác động của các số hạng bất thường nếu xuất hiện trong chuỗi thống kê. Từ đây, với mỗi giá trị quan trắc $X_o$ để kiểm tra chất lượng dựa trên thông tin khí hậu, ta xác định chỉ số $Z$ như sau:

$$Z = \frac{\left| X_0 - \overline{X}_{bi} \right|}{s_{bi}} \qquad (2.3.4)$$

Cách xác định $Z$ như vậy cho thấy phương pháp Lanzante được sử dụng tốt nhất khi $X_i$ có phân bố chuẩn như các yếu tố nhiệt độ và độ ẩm. Theo Feng và CS (2004), những quan trắc ứng với $Z > 5$ sẽ bị xem là bất thường và cần được đánh dấu nghi ngờ về mặt khí hậu. Có thể sử dụng tiêu chuẩn chặt hơn như $Z > 3$ hay $Z > 4$ nhưng theo kinh nghiệm thực tế những tiêu chuẩn này thường loại bỏ một số các quan trắc đúng. Ngoài ra, không cần thiết phải kiểm tra quá chặt ở bước kiểm tra khí hậu bởi những quan trắc sai nếu có thể qua được bước kiểm tra này còn phải qua bước kiểm tra không gian.

Thay vì xác định giá trị trung bình và độ lệch chuẩn cho từng tháng, nghiên cứu này sẽ xác định riêng cho từng ngày với mỗi trạm. Chuỗi $X_i$ sẽ được thiết lập bằng cách sử dụng mười ngày số liệu quan trắc xung quanh ngày đang tính và mở rộng sang cho mọi năm có thể. Ví dụ với ngày $N$ của năm 2009, ngoài quan trắc $X(N)$ ta đưa thêm vào chuỗi $X_i$ các quan trắc $X(N-1), ..., X(N-5)$ và $X(N+1), ..., X(N+5)$, sau đó tiếp tục mở rộng chuỗi với quan trắc từ ngày $N-5$ đến $N+5$ của năm 2008, 2007, … cho đến khi

không còn số liệu quan trắc. Bằng cách này ta sẽ có được một chuỗi đủ dài cho phép xác định giá trị khí hậu của một biến nào đó vào một ngày nhất định trong năm.

Kiểm tra khí hậu được áp dụng như đã mô tả ở trên cho các biến nhiệt độ và độ ẩm. Với mưa, kiểm tra khí hậu được kết hợp với kiểm tra vật lý. Nếu lượng mưa ngày lớn hơn 1000mm quan trắc này sẽ bị loại bỏ. Nếu lượng mưa nhỏ hơn 1000mm nhưng $Z > 5$, quan trắc bị đánh dấu nghi ngờ nhưng không bị loại bỏ mà cần được thực hiện thêm các kiểm tra khác. Các quan trắc sau khi đã qua được khâu kiểm tra khí hậu và vật lý như trên sẽ tiếp tục trải qua khâu kiểm tra phù hợp. Hệ thống thực hiện các kiểm tra phù hợp sau:

- Nhiệt độ $Tm < T < Tx$
- Độ ẩm $RHm < RH$

Nếu vi phạm các yêu cầu trên, quan trắc với yếu tố tương ứng sẽ bị loại bỏ. Cuối cùng, quan trắc ứng với mỗi yếu tố khí tượng sẽ được kiểm tra không gian theo phương pháp của Hubbard (2001). Kiểm tra không gian so sánh quan trắc tại trạm với quan trắc từ các trạm xung quanh nhằm phát hiện những bất thường có thể với số liệu đang xét. Như vậy, nguồn thông tin kiểm tra không gian dựa vào bao gồm cả các quan trắc xung quanh sẽ không loại trừ trường hợp có những quan trắc sai trong số các quan trắc này. Kiểm tra không gian hoạt động dựa trên giả thiết các quan trắc bất hợp lý xung quanh, nếu xuất hiện, chỉ chiếm một tỷ lệ nhỏ và quan trắc đang xét phải phù hợp với các quan trắc xung quanh. Định lượng sự phù hợp giữa các quan trắc này được thực hiện qua một số công cụ thống kê sẽ trình bày dưới đây dựa theo Feng và CS (2004).

Không gian các trạm xung quanh được xác định phụ thuộc vào yếu tố đang xét. Ngoại trừ áp suất, các trạm có thể đưa vào danh sách các trạm xung quanh điểm trạm đang xét cần nằm trong đường tròn với bán kính 2° quanh điểm này. Bây giờ với các trạm nằm trong đường tròn đang xét, ta xác định hệ số tương quan R giữa trạm đang xét với các trạm xung quanh dựa trên tập số liệu quan trắc quá khứ của Nd ngày, ngay trước ngày đang xét. Trong dự án này, Nd được lấy bằng 30 (tương đương một tháng). Một trạm sẽ được xem là có tương quan với trạm đang xét nếu $R >= 0.5$ và tương quan có độ tin cậy trên 95%. Tập hợp các trạm này được xác định là các trạm xung quanh của trạm đang xét.

Các trạm này sau đó được sắp xếp theo thứ tự giảm dần của R và xây dựng phương trình hồi quy tuyến tính xác định quan trắc tại trạm đang xét từ quan trắc của

mỗi trạm xung quanh. Cần chú ý rằng các quan trắc sử dụng xây dựng phương trình hồi quy đã được kiểm tra khí hậu và kiểm tra phù hợp nhằm hạn chế các số liệu quá bất thường ảnh hưởng xấu đến phương trình hồi quy. Số phương trình hồi quy sau đó sẽ được giới hạn lại bởi N = 5 phương trình, nếu có nhiều hơn năm phương trình hồi quy. Giả sử quan trắc tại trạm đang khảo sát có giá trị $X_0$. Mỗi phương trình sẽ cho ta đánh giá $X_{0j}$ với sai số $RMSE_j$ tại điểm trạm đang xét từ quan trắc $X_j$ của trạm xung quanh. Từ đây ta sẽ xác định được khoảng tin cậy của $X_0$ theo mỗi trạm xung quanh $[X_{0j} - F\times RMSE_j, X_{0j}+F\times RMSE_j]$ với F là tham số giới hạn tin cậy. Nếu $X_0$ rơi ra ngoài khoảng tin cậy với tất cả N khoảng tin cậy thu nhận được, quan trắc $X_0$ sẽ bị loại bỏ. Trong kiểm tra này, F được lấy bằng 5 với mưa, bằng 3 với các biến còn lại.

Có thể thấy phương pháp xử lý của kiểm tra không gian khá linh hoạt. Bằng cách sử dụng quan hệ thống kê qua hồi quy tuyến tính và xét tác động đồng thời của tất cả các trạm xung quanh, ngay cả khi xuất hiện những quan trắc bất hợp lý từ các trạm xung quanh, kiểm tra không gian vẫn có thể loại bỏ những quan trắc sai. Đây là một điểm mạnh của kiểm tra không gian mà các kiểm tra trước đó không có. Hơn nữa, phương pháp của Hubbart (2001) còn cho phép đánh giá tối ưu quan trắc tại trạm đang xét (khôi phục dữ liệu) từ các trạm xung quanh trong trường hợp trạm này bị mất dữ liệu. Công thức đánh giá dựa trên lý thuyết bình phương tối thiểu có dạng đơn giản như sau:

$$X_e = \sum_{j=1}^{N} \left[ X_{0j} \times RMSE_j^{-2} \right] \Big/ \sum_{j=1}^{N} RMSE_j^{-2} \qquad (2.3.5)$$

$$RMSE_e = 1 \Big/ \sum_{j=1}^{N} RMSE_j^{-2} \qquad (2.3.6)$$

Dựa theo hai công thức này, cũng có thể đưa ra khoảng tin cậy như trên và chỉ cần xét một khoảng tin cậy thay vì *N* khoảng như trên. Do đặc điểm của kiểm tra không gian không thể thực hiện độc lập riêng tại từng trạm mà phải có một số lượng nhất định quan trắc từ các trạm xung quanh tồn tại đồng thời với quan trắc đang khảo sát.