\[
\gamma' \Sigma_{12} \alpha - \mu \gamma' \Sigma_{22} \gamma = 0
\]
\[
\alpha' \Sigma_{12} \gamma - \lambda \alpha' \Sigma_{11} \alpha = 0
\]

\[\Rightarrow \lambda = \mu = \alpha' \Sigma_{12} \gamma \quad \{ \because \alpha' \Sigma_{11} \alpha = 1 \ \& \ \gamma' \Sigma_{22} \gamma = 1 \}
\]

Now from \((i)\) and \((ii)\);

\[
-\lambda \Sigma_{11} \alpha + \Sigma_{12} = 0
\]
\[
\Sigma_{21} \alpha - \lambda \Sigma_{22} \gamma = 0
\]

or

\[
\begin{pmatrix}
-\lambda \Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & -\lambda \Sigma_{22}
\end{pmatrix}
\begin{pmatrix}
\alpha \\
\gamma
\end{pmatrix}
= 0
\]

To get a non trivial solution;

\[
\begin{vmatrix}
-\lambda \Sigma_{11} & \Sigma_{12} \\
\Sigma_{21} & -\lambda \Sigma_{22}
\end{vmatrix}
= 0
\]

or \[|\Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} - \lambda^2 I| = 0\], it is a polynomial of degree

If \(\lambda^2 = \nu\) and \(\lambda = \sqrt{\nu}\) then \[|\Sigma_{22}^{-1} \Sigma_{21} \Sigma_{11}^{-1} \Sigma_{12} - \nu I| = 0\]

and if \(\nu_1, \ldots, \nu_{p_2}\) be the \(p_2\) roots of this equation, then find the maximum of

\[\sqrt{\nu_1}, \ldots, \sqrt{\nu_{p_2}}\] say \(\sqrt{\nu^{(1)}}\) and if we find \(\gamma^{(1)}_{p_2 \times 1}\) s.t.

46
\[(\sum_{22}^{-1}\sum_{21}\sum_{11}^{-1}\sum_{12} - \sqrt{V^{(1)}_I} I)Y' = 0 \quad \text{and} \quad \alpha_{pi}^{(1)} = \frac{R_{ii}^{-1}R_{iz}Y^{(1)}}{\sqrt{V^{(1)}}}\]

Then \(\alpha^{(1)}Y^{(1)} \) and \(\gamma^{(1)}Y^{(2)}\) will give the first set of canonical variates and the canonical correlations as \(\sqrt{V^{(1)}}\). In this way we will get \(p_2\) vectors of canonical weights for both \(X^{(1)}\) and \(X^{(2)}\) as \(\left[\alpha^{(1)}, \ldots, \alpha^{(p_2)}\right]\) and \(\left[\gamma^{(1)}, \ldots, \gamma^{(p_2)}\right]\) with canonical correlations as \(\sqrt{V^{(1)}}, \ldots, \sqrt{V^{(p_2)}}\).

(ii) **Multiple linear regression** (Draper and Smith, 1981)

- **Estimation of coefficients and predictand:**

  \[
  \begin{align*}
  \tilde{Y} &= \alpha_0 + \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_k x_k + \varepsilon \\
  \hat{Y} &= \alpha' X + \varepsilon
  \end{align*}
  \]

  Where \(\tilde{Y}\) is the vector of dependent variables or predictands.

and \(\tilde{x}_i (i = 1, \ldots, k)\) are the \(k\)-independent \(n \times 1\) vectors of variables or predictors.

\(\alpha_i (i = 1, \ldots, k)\) are the \(k\)-regression coefficients.

Let \(\hat{Y}\) is an estimated value of predictand and \(\alpha_i\)'s are to be estimated such that sum of squares of estimation error is minimized on the developmental sample of size \(n\), i.e.
\[ \sum_{j=1}^{n} (y_j - \hat{y}_j)^2 = \text{min} \]

An estimate of coefficients is given as;

\[ \hat{\beta} = (X'X)^{-1} X'Y \]

and the estimate of predictand would be;

\[ \hat{Y} = \hat{\beta}'X_i \]

- **Analysis of Variance:**

For testing the significance of the regression coefficients for the hypothesis;

\[ H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \]

<table>
<thead>
<tr>
<th>Source of variation</th>
<th>Degree of freedom</th>
<th>Sum of squares (SS)</th>
<th>Mean squares (MS)</th>
<th>F-statistic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Due to Regression</td>
<td>k-1</td>
<td>SS_{R} = \hat{\beta}'X'Y - n\bar{Y}^2</td>
<td>MS_{R} = SS_{R} / k - 1</td>
<td>MS_{R} / MS_{E}</td>
</tr>
<tr>
<td>Errors due to Residuals</td>
<td>n-k</td>
<td>SS_{E} = SS - SS_{R} = Y'Y - \hat{\beta}'X'Y</td>
<td>MS_{E} = SS_{E} / n - k</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>n-1</td>
<td>SS = Y'Y - n\bar{Y}^2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Where \( \hat{\alpha} = (X'X)^{-1} X'Y \)

\[
SS_R = \sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2 = \sum_{i=1}^{n} \hat{y}_i^2 - (\sum y_i)^2 / n = \hat{\Xi}'\hat{\Xi} - n\bar{Y}^2 = \hat{\alpha}'X'Y - n\bar{Y}^2
\]

As \( \hat{\Xi}'\hat{\Xi} = (X\hat{\alpha})'(X\hat{\alpha}) = \hat{\alpha}'X'X\hat{\alpha} = \hat{\alpha}'X'Y \) (\( \because X'X\hat{\alpha} = X'Y \) & \( Y' = X\hat{\alpha} \))

Now if \( \varepsilon \sim N(0, I\sigma^2) \)

Then

\[
F = \frac{MS_R}{MS_E} = \frac{\hat{\alpha}'X'Y - n\bar{Y}^2 / k - 1}{Y'Y - \hat{\alpha}'X'Y / n - k} \quad \Box \quad F(k-1, n-k)
\]

If calculated value of \( F \) say \( F_0 \) is such that

\[
F_0 \leq F_{\alpha} (k-1, n-k)
\]

Then the hypothesis is accepted that is regression obtained is not statistically significant.

and if \( F_0 > F_{\alpha} (k-1, n-k) \) then we can decide that a statistically significant regression is obtained and generally \( \alpha \) is taken as 0.1 or 0.5.

- \( R^2 \) & Analysis of Variance:

\[
R^2 = R_{y,x_1,x_2,\ldots,x_K} = \frac{SS \text{ Due to regression}}{\text{Total sum of square}}
\]

\[
= \frac{\sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2}{\sum_{j=1}^{n} (y_j - \bar{y})^2} = 1 - \frac{\sum_{j=1}^{n} (y_j - \hat{y}_j)^2}{\sum_{j=1}^{n} (y_j - \bar{y})^2} = \frac{\hat{\alpha}'X'Y - n\bar{Y}^2}{Y'Y - n\bar{Y}^2}
\]
\[ R^2 = \frac{SS \text{ Due to regression}}{SS \text{ Due to regression} + \text{residual SS}} \]

\[ = \frac{\nu_1 F}{\nu_1 F + \nu_2} \]

Where

\[ F = \frac{\nu_2 (SS \text{ Due to regression})}{\nu_1 (SS \text{ Due to residual})}; \nu_1 = k - 1 \& \nu_2 = n - k \]

\[ \frac{\nu_1 F}{\nu_2} = \frac{\hat{\alpha} X' Y - n \bar{Y}^2}{Y' Y - \hat{\alpha} X' Y} \]

& \[ R^2 \sim \beta\left(\frac{1}{2} \nu_1, \frac{1}{2} \nu_2\right) \]

- % of variance explained:

\[ R^2 = R_{y,x_1,\ldots,x_k} = R_{y,y} \text{ (Multiple correlation)} \]

\[ = \frac{SS \text{ Due to regression}}{\text{Total SS}} = \frac{\sum_{j=1}^{n} (\hat{y}_j - \bar{y})^2}{\sum_{j=1}^{n} (y_j - \bar{y})^2} \]

Where \( R^2 \) = proportion of total variation about the mean \( \bar{Y} \) explained by the Regression. It is generally expressed as a percentage multiplying by 100.

\[ \% \text{ of variance explained} = R^2 \times 100 \% \]

or \[ F = \frac{SS \text{ Due to regression} / \nu_1}{\text{residual SS} / \nu_2} \]

50
(SS Due to regression / Total sum of square)ν₂

(1 – SS Due to regression / Total sum of square)ν₁

= \frac{ν₂R^2}{ν₁(1 – R^2)}; ν₁ = k – 1 & ν₂ = n – k

If % of variance explained is more, then the value of F will be more and \( H₀ \) will be rejected.

(iii) Step wise selection procedure:

- **Procedure:-**

  - Calculate the correlation of all the predictor variables with predictand (response).

  Select the first variable to enter the regression, the one most highly correlated with the response say \( Xₐ \) and test its significance also, that is test of entry.

  - Regress \( Y \) on \( Xₐ \) and obtain the least square equation. If the overall \( F \) test shows that the regression equation is significant, then we retain \( Xₐ \), otherwise not i.e. test of exit.

  - Calculate the partial correlation coefficients of all the variables not in regression, with the response. Choose as the next variable to enter into the regression the one with the highest partial correlation coefficient say \( Xₐ \). and test its significance also for entry.

  - With \( Xₐ \) as well as \( Xₐ \) in the regression, the least square equation, \( Y = f(Xₐ, Xₐ) \) is obtained. Then each of the variable is tested for its significance and
if it is found to be insignificant then it is removed, otherwise retain in the regression i.e. the test of exit.

- Similarly select the next variable, which is having highest partial correlation say $X_c$. Then $Y = f(X_a, X_b, X_c)$ is fitted and significance of each variable is tested .......and so on.

- If the variable selected at any stage is immediately rejected, then the procedure is stopped at that stage and total number of variables selected up to previous step would give the required regression.

- **Significance level:-**

  We need two level of significance that is “exit $\alpha$” or “entry $\alpha$”. Generally these are taken to be same and we take the value of $\alpha$ as 0.5 or 0.1. But if these are taken to be different, then it will not be wise to set the “exit $\alpha$” smaller than the “entry $\alpha$”, or else one sometimes rejects predictors just admitted.

  Some experts may like to set the “exit $\alpha$” to be larger than the “entry $\alpha$” to provide some protection for predictors already admitted to the equation.

- **Stopping rule :-**

  In stepwise regression by changing value of “exit $\alpha$” or “entry $\alpha$”, we can change the number of variables finally selected and percentage of variance explained by these. But one should not take the value of $\alpha$ very seriously, instead if possible one can stop if nearly 99% of variance is explained by the finally selected variables that is unexplained variance remains only 1%. But if at any stage the percentage of additional
variance explained by the variable selected is very low that is less than 0.05 or 0.1, then also the selection of variables may be stopped.

1.2.4 Bias free rainfall and Kalman filtered temperature forecast

This is calibration technique, which is used for calibrating direct model output forecast for rainfall and maximum/minimum temperature. In this threshold values for rain/no rain for bias free rainfall forecast is based upon observed and forecasted data of last two to three seasons. The methodology for calibrating the maximum/minimum temperature forecast is based upon one parameter Kalman filter. For this, the observed and forecasted values for temperatures for last forty days is used every day and Kalman filter constants are developed and forecasts are obtained for the prognosis hours(24hr to 120hr).

The Kalman filter technology uses data from GTS surface weather parameters files archived every day. This methodology is applied and operational forecasts are obtained every day and put on website for 70 major cities of India. This forecast is based upon T-80 and T-254 models and are explained in Chapter-5.

(a) Bias free rainfall forecast.

The observed and direct model output forecast data is considered for last two to three seasons. and HK score (defined in Chapter-3) for rainfall is calculated and threshold values are made to vary and the best threshold value is chosen for the station under consideration which maximizes the HK score. Such threshold values are obtained for number of representative stations for each state and the general optimal thresholds are
obtained for each state. These optimal thresholds are used while obtaining the forecast during current season. These are explained in detail in Chapter-5.

(b) Kalman filtered temperature forecast

(i) Need for applying Kalman filter.

It is a common fact that Numerical Weather Prediction (NWP) models exhibit systematic errors in the forecasts of near surface weather parameters. This is a result not only of shortcomings in the physical parameterization but also of the inability of these models to handle successfully sub-grid phenomena. Furthermore, predictions covering areas that are not close to grid points are usually based on interpolations of the results of the models, a fact which also increases the ‘noise’ in the final outputs. The 2m-temperature, for example, is one of the most commonly biased variables, where the magnitude of this bias depends, among other factors, on the geographical location and the season. A first step towards the reduction of these types of error can be made using ‘classical’ statistical tools, such as linear regression or moving average correction methods. However, frequent changes in the versions of NWP models as well as seasonal alterations lead to analogous changes in the form of the results, a fact which, combined with the need for extended series of data, brings into question the effectiveness of traditional statistical methods.

One of the most convenient approaches to the above mentioned problems is the use of Kalman filtering (Brockwell & Davis, 1987; Kalman, 1960; Kalman and Bucy, 1961). This technique gives excellent results in the correction of systematic errors in any type of
prediction, based on the recursive combination of recent forecasts and observations. The main advantage is the easy adaptation to any changes in the data being used.

Here we present a simple (one-dimensional) Kalman filter, for the correction of systematic errors in the prediction of maximum and minimum 2m-temperatures. (Kallos et al., 1997, Homleid, 1995, Persson, 1991, and Simonsen, 1991)

(ii) A one parameter Kalman filter model.

Here we define a one-dimensional Kalman filter for the correction of systematic errors of 2m-temperature forecasts by a NWP model (Galanis and Anadraniastakis, 2002). For this, we define the measurement vector \( y_t \) as the difference between observation and the model forecast and the state vector \( x_t \) as the systematic part of the error. We work, therefore, in a one-dimensional state space. Concerning the change of \( x \) in time we have no solid evidence to rely on. As a result, we must assume this change to be random, setting the system coefficient (transition matrix) as \( F_t = 1 \). Hence, the system equation takes the form;

\[
x_t = x_{t-1} + w_t
\]

Analogously, the observation (measurement) equation is given by:

\[
y_t = x_t + v_t
\]

Obviously, \( w_t, v_t \) are here scalar variables of zero mean, since in a different case their non zero part could also be included into the systematic error.
The initial value $X_0$ of the systematic error can be assumed to be 0, unless of course we have different positive indications of its previous behavior, and the initial variance $P_0$ must, at the same time, have a considerably large value, which indicates that we do not really trust our first guess.

With the previous assumptions our Kalman filter algorithm has the following form:

- **State vector**: The space of real numbers $R$
- **State vector**: $X_i = \text{the systematic part of the error of our NWP model.}$
- **System equation**: $X_i = X_{i-1} + W_i$
- **Observation (measurement) equation**: $y_i = X_i + v_i$
- **Prediction equations**: $X_{t/t-1} = x_{t-1}$, $P_{t/t-1} = P_{t-1} + W_t$
- **Updating equations**: $x_t = x_{t/t-1} + K_t(y_t - x_{t/t-1})$, $K_t = \frac{P_{t/t-1}}{P_{t/t-1} + V_t}$, $P_t = (1 - K_t)P_{t/t-1}$

One of the most serious difficulties in Kalman filtering models concerns the way that the covariance matrices of $W_t$ and $V_t$ will be specified.

*Improved forecast for time $t+1 = (model outcome for time t+1) + filter estimate at time t)*

56
1.2.5 Bias free rainfall and trend based temperature forecast

This is also a calibration technique, which is used for calibrating direct model output forecast for rainfall and maximum/minimum temperature. In this threshold values for rain/no rain for bias free rainfall forecast is based upon observed and forecasted data of last two to three seasons. Such threshold values are obtained for number of representative stations for each state and the general optimal thresholds are obtained for each state as mentioned in the previous section. These optimal thresholds are used while obtaining the forecast during current season.

The methodology for calibrating the maximum/minimum temperature forecast is to forecast temperature trends rather than absolute temperatures. For this it can be shown that the root mean square error (RMSE) for temperature trend forecasts are always less than or equal to that for absolute temperatures. The trend based temperature forecast when tabbed for very high values of trends, gives a considerably high skill, hence such forecast can be used directly by the users. Although the forecast for the absolute temperature is to be obtained at user's end by using the current observed temperature and forecasted trend values.

In the absence of observed weather data for maximum/minimum temperature on real time basis, This methodology is applied and operational forecasts are obtained every day and put on website for 602 districts of India. This forecast is based upon T-170 and T-254 models and are explained in Chapter-6.
1.2.6 Comparative and verification studies for evaluation of skill of forecasts

In the present study the methodologies for interpretation and calibration of direct NWP models forecasts, various types of forecasts thus obtained and comparative and verification studies for the evaluation of their skill, is presented.

For these comparative and verification studies following skill scores are used, as discussed from Chapter-3 to Chapter-6 and defined in Chapter-3.

For rainfall ratio score and Hanssen and Kuiper(HK) score are used and for maximum and minimum temperature, root mean square error(RMSE) and correlation are used.

In order to have efficient comparative and verification studies these simple but critical skill scores used.

1.3 SIGNIFICANCE OF THE PROBLEM TO THE SOCIETY AND THE COUNTRY:

Location specific weather forecasts in medium range (3-5 days) are extremely useful, as 3-5 days lead time is very essential for taking precautionary measures to save crops from bad weather in agriculture and managing resources in other sectors. Such forecasts play an important role in expeditions e.g. Mountain expeditions and are used for civil and military operations both. These forecasts are also of strategic importance for the projects like satellite launching.