**­**

TONGJI UNIVERSITY

毕业设计（论文）

|  |  |
| --- | --- |
| 课题名称 | CTS-SS求解GFDL CM模式CNOP及其在ENSO事件最快增长初始误差中的应用 |
| 副 标 题 |  |
| 学 院 | 软件学院 |
| 专 业 | 软件工程 |
| 学生姓名 | 彭程 |
| 学 号 | 1352905 |
| 指导教师 | 袁时金 |
| 日 期 | 2017-06-16 |

**CTS-SS求解GFDL CM模式CNOP及其在ENSO事件最快增长初始误差中的应用**

摘 要

本文利用ENSO事件吸引子的低维性质，采用主成分分析（PCA）方法对ENSO模式进行降维，然后基于降维后的模式数据，使用Sine混沌映射生成随机初始扰动，通过连续禁忌搜索算法求解条件非线性最优扰动（CNOP），即本文选取的最快增长初始误差。

基于Geophysical Fluid Dynamics Laboratory Coupled Model（简称为GFDL CM模式）设计并完成数值实验。通过预先设定的适应度函数，设计程序实时监测模式运行结果，并将获取结果与当前最优结果进行比较，以此来不断接近最优解。最后通过对获取的最优解进行数值，图像和理论分析，判断其是否为最快增长初始误差。结果表明，计算得到的CNOP具有较大的非线性发展，其发展结构具有物理意义，有助于提高ENSO事件的可预报性。

**关键词：**CNOP，连续禁忌搜索，智能算法，GFDL CM，ENSO

**CTS-SS for Solving GFDL CM Model CNOP and Its Application on Optimally Growing Initial Errors of ENSO**

ABSTRACT

In this paper, on the property of low-dimensional attractor for ENSO, we use principal component analysis (PCA) method of ENSO model for dimensionality reduction, then use the mode of data dimensionality reduction which based on Sine chaotic map to generate random perturbations, and through continuous tabu search algorithm to solve the conditional nonlinear optimal perturbation (CNOP), which means the optimally growing initial error in this paper.

Based on the Geophysical Fluid Dynamics Laboratory Coupled Model (abbr. GFDL CM), the numerical experiments are designed and completed. Through the preset fitness function, the program is designed to monitor the running results of the model in real time, and the results are compared with the current best result to keep approaching the optimal solution. Finally, after get the obtained optimal solution, the numerical, image and the theoretical analysis are used to determine whether it is the optimally growing initial error. The result shows that the calculated CNOP has great nonlinear development, and its structure is of physical significance, which has contributed to the predictability of the ENSO.

**Key words:** CNOP,continus tabu search, intelligence algorithm, GFDL CM, ENSO

目 录

[1 引 言 1](#_Toc483756208)

[1.1 研究背景和意义 1](#_Toc483756209)

[1.2 研究现状 1](#_Toc483756210)

[1.3 研究内容 2](#_Toc483756211)

[1.4 本文结构 2](#_Toc483756212)

[2 相关研究内容 4](#_Toc483756213)

[2.1 GFDL CM模式 4](#_Toc483756214)

[2.1.1 大气模式 4](#_Toc483756215)

[2.1.2 海洋模式 4](#_Toc483756216)

[2.2 智能算法求解CNOP 5](#_Toc483756217)

[2.2.1 CNOP定义 5](#_Toc483756218)

[2.2.2 智能算法综述 6](#_Toc483756219)

[2.2.3 特征提取方法 7](#_Toc483756220)

[2.2.4 智能算法求解CNOP流程 8](#_Toc483756221)

[2.3 ENSO事件最快增长初始误差 9](#_Toc483756222)

[3 GFDL CM模式样本数据集分析和降维研究 12](#_Toc483756223)

[3.1 GFDL CM模式样本数据集 12](#_Toc483756224)

[3.1.1 NetCDF文件格式说明 12](#_Toc483756225)

[3.1.2 GFDL CM模式数据采集 12](#_Toc483756226)

[3.2 PCA方法提取主特征 14](#_Toc483756227)

[3.3 数值实验与分析 14](#_Toc483756228)

[4 CTS-SS求解CNOP 17](#_Toc483756229)

[4.1 禁忌搜索算法 17](#_Toc483756230)

[4.1.1 基本原理 17](#_Toc483756231)

[4.1.2 算法流程 18](#_Toc483756232)

[4.2 改进的禁忌搜索算法 19](#_Toc483756233)

[4.2.1 连续禁忌搜索 19](#_Toc483756234)

[4.2.2 基于Sine混沌映射生成初始值 20](#_Toc483756235)

[4.2.3 分段式搜索策略 23](#_Toc483756236)

[4.3 CTS-SS求解CNOP算法描述 23](#_Toc483756237)

[5 实验与验证 25](#_Toc483756238)

[5.1 实验环境 25](#_Toc483756239)

[5.2 程序说明 25](#_Toc483756240)

[5.3 实验步骤 26](#_Toc483756241)

[5.2.1 获取SST数据及降维 27](#_Toc483756242)

[5.2.2 获取最优初始解 28](#_Toc483756243)

[5.3.3 CTS-SS寻优与实验结果对比 31](#_Toc483756244)

[5.3.4 得到最优解并输出 31](#_Toc483756245)

[5.4 实验结果分析 31](#_Toc483756246)

[6 结论和展望 39](#_Toc483756247)

[6.1论文总结 39](#_Toc483756248)

[6.2 进一步研究方向 39](#_Toc483756249)

[参考文献 40](#_Toc483756250)

[谢辞 41](#_Toc483756251)

# 1 引 言

## 1.1 研究背景和意义

大气和海洋科学中的数值天气和气候可预报性研究一直是国内外研究的热点问题，目前已经有很多方法被建立及应用于探讨数值天气和气候的可预报性问题。由于初始扰动的选择和发展会对预报结果的准确性和稳定性产生重要影响，因此，如何准确描述气候发展变化的非线性系统中的最快增长初始扰动成为上述问题中一个重要的研究领域。

穆穆等人提出了条件非线性最优扰动（Conditional nonlinear optimization perturbation, CNOP）[1]的概念，它是满足一定约束条件且在预报时刻具有最大非线性发展的一类扰动。CNOP方法充分考虑了动力系统的非线性特征，衡量了非线性在动力系统从初始时刻演变到预报时刻过程中起到的作用，为研究大气和海洋科学的可预报性问题提供了有力的工具。目前CNOP方法已经在大气和海洋科学中取得了广泛的应用。

天气与气候的发生与演变过程可以用一组非线性偏微分方程组加以描述，针对所关注的各种问题，从物理角度对方程组进行简化，例如，通过尺度分析、地转近似和静力平衡近似等得到不同的简化方程组，这就是各种“模式”，在进一步将这些模式转化为计算机可以数值求解的数值问题，就是通常意义上的数值模式。结合CNOP方法，使用数值模式提高各类气候事件（如ENSO事件）的可预报性，是当代短期气候预测研究中的一个重要课题。

ENSO是厄尔尼诺（El Niño）和南方涛动（Southern Oscillation）的合称，是短期气候年际变化的最强信号之一。ENSO现象是引起全球大气环流和水分循环异常的重要原因，全球各地许多灾害性的气候事件都与其相关。科学研究表明，ENSO是导致全球各地破坏性干旱、暴风雨和洪水的罪魁祸首。1997至1998年的厄尔尼诺导致成千上万人死亡，并造成世界直接经济损失达数十亿美元。同时，我国发生的98年长江洪灾、南方雪灾等都有其有关，所以其可预报性研究一直是学术界关注的热点问题，CNOP方法更是首先被应用于ENSO事件的可预报性研究中[2]。

综上，本文选取Geophysical Fluid Dynamics Laboratory Coupled Model（简称GFDL CM模式）作为数值模式，运用CNOP方法寻找ENSO事件中具有实际物理意义的最快增长初始误差（Optimally growing initial errors，简称OGE），并将其应用于ENSO事件的可预报性问题研究中，其结果对ENSO的可预报性研究具有非常重要的科学意义。

## 1.2 研究现状

求解CNOP的过程本质上是获得最优解的过程，而最优解的定义取决于具体求解的问题。在ENSO事件可预报性研究中，最快增长初始误差是指发展最快且导致不容忽视的预报误差的初始误差。作为ENSO事件可预报研究的一个重要内容[3]，研究OGE的发展变化可以帮助我们更好地理解ENSO事件的物理机制和提高ENSO的预报技巧，寻找ENSO事件中OGE就是本文所需要寻找的最优解。目前，伴随方法是求解CNOP最流行的方法，但使用伴随方法完成一次数值实验往往需要耗费大量的时间，而获得一个有意义的结论又需要通过大量的数值实验[4]。尤其对于复杂的全球海气耦合模式（如本文所采用的GFDL CM模式）而言，开发其相应的伴随模式往往需要耗费数年时间，工作量巨大，目前还没有开发出基于GFDL CM模式的伴随模式。但从求解问题的角度而言，伴随模式只是计算梯度的一个有效工具，并不是必不可少的步骤，作为计算机领域最优化问题求解的一种有力工具，智能算法提供了一种新的求解思路。

智能算法是人们受到自然规律的启发而建立的求解问题的一类算法，其目的是解决最优化问题。由于其求解过程不依赖梯度信息，可以用于求解CNOP这类免伴随的问题。根据搜索个体的数目，智能算法大致可以分为单粒子智能算法和多粒子智能算法。温仕成[5]比较了多种智能算法在求解CNOP过程中结果的正确性和稳定性。实验表明，无论是单粒子智能算法还是多粒子智能算法，通过优化智能算法求解CNOP都可以使其能够应用于中等复杂程度以上的数值模式，这为本文的研究提供了理论依据。

## 1.3 研究内容

本文主要研究利用基于Sine映射和分段式搜索的连续禁忌搜索算法（Continuous Tabu Search Algorithm with Sine Maps and Staged Strategy，简称CTS-SS）[6]求解GFDL CM模式CNOP，并设计数值实验，将其应用于寻找ENSO事件中的OGE之中。主要进行的工作有：

（1）研究模式数据的特征提取。GFDL CM作为全球海气耦合模式，直接应用智能算法求解必将导致收敛速度过慢等问题，使用特征提取方法对原始数据进行降维是对其应用科学计算的前提条件。通过MATLAB等数学工具，使用主成分分析（Principal component analysis，简称PCA）方法对模式数据的初始值进行降维处理，同时比较不同的PCs对最终结果的准确性和稳定性产生的影响。

（2）设计并调整CTS-SS算法参数。由于CTS-SS仅仅是一种算法思想，在实际使用过程中，我们需要结合具体问题确定其中各项具体参数，其中包括邻域的设置，交换过程的定义和禁忌判断参数的大小等，都需要基于数值实验的分析和调整。尤其是面对GFDL CM这种高复杂度的海气耦合模式，初始值的不同会对实验结果造成非常大的影响。

（3）设计并完成数值实验。对于寻找具有实际物理意义的OGE，如何设置合适和目标函数和合适的约束条件是设计实验时需要重点考虑的问题。这部分工作中，我们实现了完整程序代码（基于Java编程语言），将调整过后的CTS-SS算法应用于GFDL CM模式动态求解（每次模式运行结束立即即获取实验结果，并进行比较分析），不间断地轮询模式运行状态，将所得结果及时作为输入参数应用于CTS-SS各个阶段的寻优过程。并结合实际理论，运用图像，数值分析等手段，验证输出结果的正确性和稳定性。

## 1.4 本文结构

根据本文的研究内容，论文的组织结构安排如下：

第一章 引言。主要介绍本文的研究背景和意义，阐述本文的主要研究内容以及具体使用的研究方法和理由。

第二章 相关研究内容。主要介绍本文选用的GFDL CM模式，描述其各个子模块的关系和选择理由。之后介绍智能算法求解CNOP的应用，并分析其优缺点。最后介绍本文实际求解的问题，即ENSO事件中的最快增长初始误差。

第三章 GFDL CM模式样本数据集分析和降维研究。阐述了GFDL CM模式数据的选择，具体的降维方法及过程，重点描述PCA方法在GFDL CM模式数据的具体应用，同时设计实验以比较不同PCs的选择对实验效率和结果造成的影响。

第四章 CTS-SS求解CNOP。本章主要介绍CTS-SS的基本理念和在实际求解GFDL CM模式CNOP时需要进行的结构和参数调整。

第五章 实验与验证。本章主要介绍整体实验的设计思路和具体实现，描述了数值实验的具体运行原理和流程，并对实验结果进行分析和讨论，并针对一个较好的实验结果进行了全面的数值和图像分析。

第六章 结论和展望。对上述工作进行总结，并对今后的研究进行了展望。

# 2 相关研究内容

## 2.1 GFDL CM模式

气候模式是用于提高人们对各种时间尺度（季节，年度，年代等）气候现象的理解和可预报性的重要工具，其预测结果能为人们各类决策提供必要的信息，如水资源管理，农业，交通和城市规划等。从本质上来说，气候模式是用于描述气候系统中的主要成分（大气，陆地，海洋和海冰等）动态发展及其相互作用的一种数学模型。其中，GFDL CM模式是由美国国家海洋大气管理局地球流体动力学实验室(Geophysical Fluid Dynamics Laboratory)开发的一个复杂全球海-气耦合环流模式。其始于20世纪60年代，经过50余年的发展，目前已经在气候变化、 热带气旋等的研究中得到广泛的应用。目前GFDL CM模式已被证明在ENSO事件可预报性问题中具有较高的可信度[7]，这也是本文选取其作为运行模式的原因。

本文所采用的GFDL CM模式版本为GFDL CM2p1，该模式共由大气、海洋、陆地和海冰四个子模式组成，其中最重要的两个子模式为大气模式和海洋模式，下面分别介绍这两个子模式。

### 2.1.1 大气模式

GFDL CM2p1大气模式垂直方向上共24层， 利用时间步长为3小时的大气辐射和时间步长为0.5小时的其他大气物理量进行模式积分，并包括一个日照昼夜循环。其水平分辨率2°\*2.5°。该分辨率下足以分析产生大中型气旋等天气变化的原因。

### 2.1.2 海洋模式

GFDL CM2p1海洋模式垂直方向上共50层，其水平分辨率在经度方向为1°，在纬度方向由极地地区的1°变化到赤道地区的1/3°。对于垂直方向，其最高22层中每层厚度为10米，共220米，该分辨率下可以比较清楚地辨别ENSO事件的发生与否。同时模式包含一个贯穿欧亚大陆、北美洲和南极洲的三极网络格点，以避免北极上空出现极性滤波，其中大气和海洋模式每两小时进行一次分量的通量交换。图2.1为使用控制实验数据（51~151年），运行GFDL CM2p1模式生成的海表面温度距平（Sea Surface Temperature Anomaly，简称SSTA）图像。从图像可以看出，GFDL CM2p1在图像精确度和生成结果稳定性上，均具有良好的效果。



图2.1 GFDL CM2p1 模拟SSTA全球分布图

## 2.2 智能算法求解CNOP

### 2.2.1 CNOP定义

CNOP是指满足一定的约束条件下，在预报时刻会对预报结果不确定性产生最大影响的一类初始扰动。就其本身求解过程而言，首先需要考虑如下初值问题：

 （2.1）

其中*N*是非线性算子，*U*用来表示海表面温度异常、斜温层深度距平和表层流速距平，*U0*表示*U*的初始向量。方程（2.1）在*τ*时刻的解可以表示为：

 （2.2）

其中，假设*Mτ*表示从初始时刻到τ时刻的非线性传播算子,*u*0表示初值为*U0*的基态的初始扰动，将这一初始扰动叠加于状态向量*U0*，可以得到：

 （2.3）

其中，代表了初始扰动*u*0 的非线性发展。

CNOP表示这样的一类初始扰动，其满足目标函数在给定初始约束条件（约定，具体约束条件取决于具体的问题）时取到最大值，即：

 （2.4）

其中，

 （2.5）

*U0*代表基态的初态场，*u*0表示初始扰动。为了方便问题求解，我们通常把该最大值问题转化为求解如下的最小值问题：

 （2.6）

### 2.2.2 智能算法综述

智能算法是指人们受到自然（生物界）规律的启发，模仿其求解问题的一类算法，也被称为现代启发式（Modern heuristics）方法或计算智能（Computational intelligence）。这类算法可以上溯到最优化技术，也就是研究在众多解决方案中什么样的方案是最优的，或者怎样找出最优的解决方案。

现实世界中的许多工程问题或管理问题都可以归结为带约束的最优化问题，其中约束和最优的定义都依赖于具体求解的问题。总体说来，最优化问题可以分为函数优化问题和组合优化问题两大类，其中，函数优化问题是指求解对象为一定区间内的连续变量，而组合优化问题是指求解对象为解空间中的离散对象。

对于经典的数值方法（如解析法，数值法）而言，其求解上述两类最优化问题过程中最大的困难在于目标函数通常存在诸多局部最优解，而我们的数值方法通常是在局部区域内求得最优解，而不是获得整体解空间上的最优解[8]。尤其是对于组合优化问题，其计算复杂度很高，属于非确定性多项式困难问题（non-deterministic polynomial-time hardproblem，NP-hard），除了通过枚举法穷尽一部分解空间外，没有更好的解法。同时随着问题规模的增大，解空间会呈现出指数级甚至阶乘级增长，这个时候希望求得准确的最优解实际上已经不可能。而智能算法便是相对于最优算法提出的，一个问题最优算法可以求得该问题的最优解，而智能算法则可以认为是一种基于直观或经验构造的算法，通过可以承受的花费（时间复杂度和空间复杂度等）来获得待解决优化问题的一个可行解，该可行解与最优解的偏移程度不一定可以预先估计，但从工程实际的角度来看，在一定误差范围内的近似解都是可以接受的。同时，由于数学模型（比如本文主要研究的CNOP方法）本身就是对实际问题的简化，或多或少会忽略一些因素，加之数据采集和参数估计准确性等原因，可能会使最优算法所求得的解与智能算法获得的解产生一定误差，这是应用智能算法求解时需要尽量回避的问题。

由于智能算法速度快，操作相对简单等特点，其已经得到了很快的发展。无论于工程中的许多问题，比如通讯网络的结构优化，管道重组等，还是计算机科学中的许多问题，如旅行商问题、0-1背包问题、聚类问题等都获得了广泛的应用。本课题组当前已经将智能算法广泛应用于各类CNOP求解过程中，并取得了一定的研究成果[9-10]，但目前还没有将智能算法应用于求解GFDL CM模式中CNOP的工作，这也是本文的可行性和创新性。

### 2.2.3 特征提取方法

对于中等程度以上的模式来说（比如本文采用的GFDL CM模式），其问题的数据规模一般在106  以上，而智能算法一般只能直接应用于上述的小规模问题。当问题规模较小时，智能算法能够很快的求出全局最优解。然而，随着问题规模的增大，智能算法需要增加个体规模来获得对全局系统的搜索，这会导致收敛速度的锐减，甚至造成不收敛的情况，这个时候就需要预先对数据进行降维处理。将数据从原始的千万级别的高维空间转化为一个数十维的低维空间，之后再在该低维空间上应用智能算法求解CNOP。特征提取方法求解高维度问题的有效性已被文献[11]所证明，同时，Zhang等[10]已经将PCA作为特征提取方法，运用于智能算法求解CNOP，并取得了理想的结果。下面以PCA为例，介绍特征提取方法的主要原理：

假设由*n*个历史样本*x1，x2，...xn，xi∈Rm（i=1，2...n）*组成一个数据样本矩阵。首先对该矩阵进行无量纲化和中心化两步预处理。无量纲化是指将有量纲的物理量（比如SST等）通过数值方法来避免量纲不统一引起的数值计算的问题，其具体操作如下式：

 （2.7）

其中*a*是一个正常数。

中心化是指每个样本减去样本均值，操作如下式：

 （2.8）

这样样本数据就变成了*x’i*（*i*=1, 2...*n*），新生成的样本矩阵为*X*’=（*x’*1, *x’2*...*x’n*）。经过数据预处理，我们使用PCA方法对数据进行降维，其目的是为了求得这样一组特征向量*β*i，使得*x’i（i=1，2...n）*在其方向的投影最大，用公式表述如下：

 （2.9）

经过验证可知，和分别为样本矩阵的协方差矩阵*M*的特征值和对应的特征向量。

 （2.10）

其中是一个的矩阵，*M*是一个的矩阵。通常情况下，，因此，若直接对矩阵*M*进行特征分解，计算量非常大（实验表明，在MATLAB等软件上运行甚至会因运算量过大无法直接求得结果）。但是，矩阵是一个，若对它进行特征分解，计算量会小很多。其特征分解过程为：

 （2.11）

和分别为矩阵的特征值和特征向量。对上式两端分别乘以可得：

 （2.12）

是就协方差矩阵*M*对应特征值的特征向量，因此，通过这种方式间接地求出了特征基向量：

 （2.13）

因为样本数据已进行了中心化处理，所以其协方差矩阵的秩不超过（*n*为样本个数）。

### 2.2.4 智能算法求解CNOP流程

基于对智能算法理论和可行性的研究，结合具体的CNOP方法求解过程，我们可以梳理出智能算法求解CNOP流程，如下图所示：



图2.2 智能算法求解CNOP流程图

根据流程图，具体阐述智能算法求解CNOP流程：首先，输入原始数据集，对数据进行上述无量纲化和中心化预处理步骤，得到样本矩阵*X’*，之后通过公式（2.13）得到特征向量*β*1、*β*2、... *β*k，将其组成主成分矩阵*C*，即*C*=[*β*1, *β*2, ... *β*k]，之后将初始扰动投影至该主成分空间，即*u0=C·ω*，在这里*ω*为初始扰动*u0*投影后的坐标。这样一来，公式（2.6）可以转换为：

 (2.14)

此时的自变量从*u0*换成了*ω*，由于主成分矩阵C的关系，*ω*的维度为k维而不再是原先*u0*的m维，且k << m。事实上，公式（2.14）的约束条件||C·*ω*||≤*δ*，可以被进一步如下简化：

 (2.15)

此时，CNOP问题的适应度函数就可以借助公式（2.15）转化为：

 (2.16)

公式（2.16）就是在提取主特征、完成数据降维后交由智能算法计算的适应度函数，即最优化问题中度量。初始化个体即初始化*ω*。

需要注意的是，特征空间相当于是个超球体，所以在特征空间中应用智能算法更新*ω*，往往会使得*ω*的位置移动到超球体之外，因此需要用如下投影公式，将被投射到边界之外的点在重新投影回特征空间：

 (2.17)

公式(2.17)即为投影公式，式中*δ*即为特征空间半径。

## 2.3 ENSO事件最快增长初始误差

ENSO是厄尔尼诺（El Niño）和南方涛动（Southern Oscillation）的合称，是发生在横跨赤道附近太平洋的一种准周期气候类型。其中，El Niño发生在海洋里，是指赤道东太平洋海水大范围的异常增暖现象。根据美国海洋与大气局（NOAA）的定义—如果Niño3.4区域（170°W~120°W，5°S~5°S）的SST值与1971年到2000年的气候态平均之差连续3个月大于或等于0.5°C时，可以认为发生了一次厄尔尼诺事件，若距平连续3个月小于-0.5°C时，则可以认为发生了一次拉尼娜事件（图2.3代表了国际上主流动力预报模式在2017年5月对NINO3.4区域SSTA的预报）。而南方涛动现象发生在大气中，指的是南太平洋高压和印尼抵押之间海平面起亚的震荡，这两种变动是相互联系的。



图2.3 国际上主流动力预报模式在2017年5月对NINO3.4区域SSTA的预报

 图2.4代表一次较为典型的ENSO事件海表面异常分布（1998年）。随着人们对ENSO物理机制的不断深入了解，目前已经有很多方法用于提高ENSO的预报技巧。

图2.4 典型ENSO事件SSTA分布

在文献[3]中，陈磊已经指出ENSO预报的不确定性来自于气候平均态的年循环、ENSO事件本身和初始误差场的结构这三个因素。在本文中，我们采用完美模式假设，即仅考虑初始误差场的结构这一因素。

最快增长初始误差是指发展最快且导致不容忽视的预报误差的一类初始误差。最快增长初始误差可分为两类：一类为局部条件非线性最优扰动型初始误差，称为type-1型最快增长初始误差，其主要表现为扰动在热带东太平洋从次表层到表层符号一致的自西向东倾斜结构；另一类为全局条件非线性最优扰动型初始误差，称为type-2型最快增长初始误差，其表现为扰动在热带太平洋次表层时的偶极子结构。

对于本文所使用的GFDL CM模式而言，由于其属于全球海气耦合模式，误差发展是全球发展型的，所以我们主要基于type-2型最快增长初始误差设计数值实验。作为以应用计算机科学求解地球物理科学相关实际问题的课题，本文主要关注智能算法的寻优过程，故在此处仅简要介绍ENSO事件最快增长初始误差。总体说来，为了研究ENSO的最快增长初始误差，需要预先确定作为参考态的El Niño事件，并基于此来进行数值实验，判断某次事件是否为最快增长初始误差还需要借助理论手段的分析过程。实现细节部分将在第5章详细介绍。

# 3 GFDL CM模式样本数据集分析和降维研究

## 3.1 GFDL CM模式样本数据集

### 3.1.1 NetCDF文件格式说明

NetCDF是网络通用数据格式（Network common data form）的缩写，是一种面向数组型并适用于网络共享的数据描述和编码标准，由于其自描述性、易用性、平台无关性等特点，在大气科学、海洋学、地球物理等学科中被广泛使用。GFDL CM2p1中所有的数据文件格式均为NetCDF格式，这里首先对NetCDF做简要说明。

从数学角度来说，NetCDF存储的就是一种多自变量的单值函数，可以用公式表示为*f(x,y,z,...)=value*，函数的自变量x，y，z等在NetCDF中称为维度或坐标轴，函数值value在NetCDF中称为变量，而自变量和函数值在物理学上的一些性质，比如计量单位、物理学名称等等在NetCDF中称为属性。NetCDF的标准格式如下：

NetCDF name{

Dimensions：… //定义维数

Variables：… //定义变量

Attributes：… //属性

Data：…//数据

}

NetCDF的默认文件后缀名为.nc， 调用此类文件前需要进行一定的预处理，可结合具体使用数学工具或编程语言实现调用。

### 3.1.2 GFDL CM模式数据采集

 对于本文所选用的GFDL CM2p1模式来说，其主程序采用Fortran语言编写，输入文件独立于主运行程序，全部存储在/INPUT文件夹中，模式运行流程图如3.1所示：

图3.1 GFDL CM2p1模式运行流程图

模式结束运行后，运行结果会存储在/history文件夹下，其包含atmos\_month.nc和ocean\_month.nc两个文件，分别用于存储大气和海洋的月度相关数据。对于本文重点考虑的厄尔尼诺事件来说，我们仅需考虑海洋相关数据。对应输入数据文件中与海洋相关的输入文件结构如3.2所示：



图3.2 海洋模式相关输入文件

作为全球海气耦合模式，理论上来说，在上图中任何海洋相关输入文件中加入扰动，均会对实验结果产生影响。但理论分析和实验结果均表明，对于一些与SST直接关系不强的因素（如风向，海风速率等）而言，在这些物理量上即使加入较大初始扰动，对最终结果的生成也不会有较大影响（甚至可以忽略不计），故本文仅考虑在与SST直接相关的ocean\_temp\_salt.res.nc文件中加入初始随机扰动，并取其中z轴的第一层（海表层）。根据模式的分辨率，数据初始维度为360\*200\*1（xaxis\_1\* yaxis\_1\* zaxis\_1）。

## 3.2 PCA方法提取主特征

在第二章已经提过，在进行智能算法寻优前，需要对模式数据进行降维处理。然而，进行特征提取的方式有很多（如PCA,ONPP等），每一种都有其应用范围和局限。文献[5]中对比了四种不同的特征提取算法（PCA、RPCA、ONPP和OLPP），结论指出作为一种用于提取数据线性特征的降维方法，PCA提取出的特征具有良好的统计意义，在样本数据误差满足小、独立分布的条件时，可以得到理想的结果。

这里结合GFDL CM模式，进一步分析进行寻优步骤前，需要使用PCA进行降维处理的两个原因：一方面，智能算法（如本文选用的CTS-SS）是在主特征空间上寻优，所以其提取的主特征必须是相互正交的，这可以确保消除各维度数据间的相关影响，获取更客观的结果。在GFDL CM模式中，各级输入数据都被独立保存，互不影响，这保证了应用PCA的先决条件；另一方面，智能算法求解过程要求CNOP能被主特征线性表示，若选用非线性降维方法提取主特征，则CNOP不能被线性表示。基于以上两个原因，本文选取PCA方法进行数据降维，下面介绍在GFDL CM2p1中的具体做法。

首先利用GFDL CM模式积分足够长的时间（100年）得到一个训练样本集。假设模式的历史数据样本为，其中，*hi*（）由*v*个模式变量组成，且每个变量是一个的矩阵，*row*和*col*分别为模式的网格的行数和列数（在GFDL CM中分别为360和200）。现把模式变量的矩阵以列为单位串起来，变为向量形式，并将所有变量串成一个更长的向量，。样本集转变为， ，并且根据式（2.7）进行无量纲化。这样我们得到了一个72000\*100的高维矩阵，之后可在此矩阵进行PCA处理。

## 3.3 数值实验与分析

使用PCA进行主成分分析时，主成分个数的选择是一个值得探讨的问题。这个问题在不同的研究领域有不同标准，并没有确切的答案。例如在图像处理领域，大多认为如果最大的k个特征值占整个特征值的90%以上，那么取这k个特征值对应的特征向量即可。文献[13]指出，PCA的选择取决于具体求解的问题，多数情况下可取70%~90%。下面我们以GFDL CM模式数据为例，讨论主特征数目的选取。



图3.3 主成分累计特征比例分析，柱状图和折线图分别表示相应主成分下的累计特征比例

通过PCA算法，计算GFDL CM 模式在ENSO事件最快增长初始误差的样本数据集中最大的前100个特征值，其结果如图3.3所示。X轴表示主特征的数目，Y轴表示当前特征值占所有特征值的比例，简称为累计特征值比例，用于表示主特征所保持的能量。从图中可以看出，当取前10个主特征时，累计特征值比例已经超过40%，在主特征小于40之前，累计特征值比例都保持较快的增长，在主特征值大于80之后，累计特征值增长较为缓慢，基本趋于平稳。可以认为这时再增加的主特征中，其包含的原始数据量已经微乎其微了。

对于数据降维而言，除了考虑特征值分布，还需要分析降维引起的数据误差。经过降维后，原始的协方差矩阵变为：

 （3.1）

其中，由最大的*k*个特征值对应的特征向量组成的矩阵，*Dk*为一个对角阵，对角线上的元素为最大的*k*个特征值。援引文献[5]的做法，可以通过下式来评价取*k*个主特征向量所引起的相对误差：

 （3.2）

其中，是F范数，表示矩阵所有元素的平方和再开方。利用上式考察取不同主特征数目的情况下所引起误差的大小，其结果如图3.4所示，当只取10个主特征向量时，恢复得到的协方差矩阵的相对误差达到0.07；随着主成分数目的增加，相对误差会逐渐减少；当取40个主特征向量时，相对误差约只有0.01；当取到80个主特征向量时，相对误差仅为0.005左右，这时利用主特征恢复的协方差矩阵已经比较接近原始的协方差矩阵了。

 图3.4 相对误差（X轴表示主特征数目，Y轴表示相对误差。）

因此，通过对累计特征值比例的分析和对协方差矩阵恢复引起的相对误差的分析，可以看出随着主特征数目的增加，保存原始数据信息越多。但是随着主特征数目的增加，势必增加程序的计算开销。同时，随着主特征数目增多，智能算法计算出的CNOP未必会更好[5]。所以在考虑主特征数目时还需进一步结合时间开销、CNOP值和数值模态等来确定，具体在第五章实验设计时介绍。

# 4 CTS-SS求解CNOP

## 4.1 禁忌搜索算法

禁忌搜索（Tabu-Search）是最早由Glover[13,14]提出的一种全局迭代寻优算法，其主要思想是通过引入一个灵活的存出结果和相应的禁忌准则来避免迂回搜索，从而保证搜索的有效性和全局性。同时，TS也是人工智能的体现，其最重要的思想是标记那些已经寻找过的局部最优对象，并在之后的迭代中尽量避开这些对象，但并没有绝对禁止这些对象，而是通过灵活的数据结构来保证对解空间中的不同区域进行有效搜索。

作为单粒子的智能算法，TS相较于依靠大量粒子进行大范围搜索的群智能算法（例如PSO等）来说，其特点在于其由于只有一个个体，整体的寻优结果都取决于一个粒子，所以其更新规则需要合理而缜密的设计，否则粒子方向或步长的不合理设置会导致很大的误差甚至错误的结果。同时在每一步迭代中，个体更新需要的时间复杂度和空间复杂度都较多粒子算法更高。

### 4.1.1 基本原理

禁忌搜索算法中包含邻域（neighborhood）、禁忌表（tabu list）、禁忌长度（tabu length）、终止规则（termination criterion）等基本概念，这些概念是算法设计的关键，下面分别介绍这些基本概念。

1. 邻域

邻域是TS算法中最基本的一个概念，但其在具体问题中的定义却各不相同。在一些离散问题（如Traveling salesman problem）中，邻域的定义通常为进行一次数据交换后的结果，其本质是当前解的一次选择状态变化；而在距离空间中，邻域的定义则是以一个点为中心的圆。总体说来，领域可以理解为当前粒子下一步可以达到的状态。4.3将结合CTS-SS说明CNOP求解过程中邻域的定义。

1. 禁忌表

为了避免重复工作和可能陷入的局步内循环，TS建立并维护一张禁忌表以禁忌一定时期内进行的工作。其中，禁忌对象和禁忌常数是设置禁忌表时需要考虑的两个重要指标。

1. 禁忌对象

文献[15]将禁忌对象分为三类：简单解，向量分量变化和目标函数值。其分别对应不同的禁忌规则。

对于简单解来说，当解发生一次移动（比如从*x0*移动到*x1*）时，*x1*可能是当前局部最优解，为了避开局部最优解，算法将禁忌*x1*。其禁忌规则是：当*x1*的邻域中有比它更优的解时，则选择最优的解；若*x1*为当前解空间（或本次迭代）的局部最优解时，不再选择*x1*，而选择*x1*以外的其他最优解。

向量分量变化是指向量分量的相互交换。例如，考虑一个5城市的旅行商问题（记其城市编号为A~E），设当前解由*x0* =（ABCDE）变化到*x1* =（ACBDE），其变化是由B和C的对换引起的，而实际上B和C的对换可以引起更多解的变化。比如：（ABDCE）->(ACDBE),(AEDBC)->(AEDCB)等等，这时如果我们将B和C的交换禁忌，那么这些情况将都被禁忌。

目标函数值是求解最优化问题（如CNOP求解）中直接衡量实验结果的度量，而禁忌目标函数值的定义如同等位线的道理一样，将处在同一位置的一视同仁。一旦一个目标函数值被禁忌，则那些有此目标函数值得解将都被禁忌。

1. 禁忌长度

禁忌长度是指被禁忌对象不再被选取的迭代次数，具体实现可以给被禁忌对象赋值（禁忌长度）*L*，要求对象*x*在*L*步迭代内被禁，禁忌表采用tabu（x）=*L*来存储禁忌关系。每迭代一步，所有被禁忌对象就向前一步，具体实现时其可以设计为一个队列。值得一提的是，禁忌表的长度选择并没有非常固定的衡量标准，其可以设置为常量，亦可设置为约束条件下的变量。其参数的选择将在第五章实验设计时说明。

1. 终止规则

作为一个启发式算法，TS不可能让寻优过程无限进行下去，只希望在可接受的时间内获得一个相对满意的解，所以终止规则的设定也是一个需要考虑的问题。一般而言，具体问题中有如下几种终止规则：

（1）确定终止步数。给定一个合理且充分大的数*N*，算法总的迭代步数不超过*N*步。这

种做法不考虑算法中其他包含的终止原则，可以保证算法总的迭代次数。其优点是易于操作和可以控制和计算实验时间，缺点是无法保证解的效果。

（2）控制计算频率。当某一个解，其目标函数值得频率超过一定标准时，如果对算法不

加以改进，只会造成该频率的增加，这时候继续进行循环已经不会对解有改进，应当终止计算，即如果不改进算法，解将无法改进。

（3）目标函数值变化控制原则。在TS搜索中，可以采用记忆当前最优解的方法，如果在一定步数内，目标函数值没有变化，宜采用与（2）相似的策略，认为解不再改进，停止运算。

### 4.1.2 算法流程

基于以上的算法要素，可以归纳出禁忌搜索智能算法的一般步骤：

步骤1：随机生成一个初始点*x0* ，计算其目标函数值，同时初始化当前解*x*= *x0*，最优解*xbest*= *x0*，*f(xbest)=f(x0)*。

步骤2：生成当前点*x*的邻域，计算邻域内各点的目标函数值。

步骤3：根据适应度函数，选择邻域目标函数值最佳的点。

步骤4：判断是否被禁忌。如果没有，新的当前解移动到，转到步骤5，否则将从邻域中删除，回到步骤3。

步骤5：更新禁忌表，并判断终止规则。若终止规则满足，则终止计算，否则会到步骤2。

以上为传统的禁忌搜索算法流程，下面根据实际的CNOP求解问题，介绍改进的禁忌搜索算法及其具体应用。

## 4.2 改进的禁忌搜索算法

### 4.2.1 连续禁忌搜索

TS是基于组合优化问题而提出的，其在组合优化中的研究已经相当广泛，并取得了一系列研究成果，有学者尝试将其应用于连续优化问题领域的求解过程中，并由此提出了连续禁忌搜索(Continuous Tabu Search，简称CTS)的概念。需要注意的是，从求解组合优化问题到连续优化问题，4.1中所提及的TS基本思想均需要做一定的改进。本文参考文献[16]及GFDL CM模式的实际需要，对CTS中各因素的改进做如下说明：

1. 初始解的生成

 在解空间中，获得初始解的常用方法是使用随机函数，这可以在一定程度上保证解的全局性（随机函数结果如图4.1）。然而，CTS作为单粒子算法，其初始值的选择会直接影响迭代过程的效率，也会影响模式结果。我们在CTS-SS中引进Sine映射方法生成初始解，这一部分在4.2.2中详细介绍。

图4.1随机函数生成初始值散点图

1. 邻域的生成

如4.1.1所述，邻域的定义通常取决于具体求解的问题。在求解多维问题的CTS算法中，邻域被定义为一个高维超球体的区域，对应于一个当前解，其邻域内有无数个点来进行下一步交换。出于计算效率的考量，设定超球体邻域的半径之后，在数值实验中随机选取一定数目的解作为候选解（具体实验在第五章介绍），同时结合分段式搜索策略，将算法的搜索过程分为多个阶段，每个阶段的不同之处在于候选解的个数和超球体邻域的半径值。其搜索思路是：在第一阶段，由于初始解可能远离最优解，尽量选取较大的邻域半径和较少的候选解。随着搜索的进行，实验结果越来越逼近最优解，这时选取较小的邻域半径和较多的候选解，进行局部集中搜索。整体搜索由扩散到聚焦，其详细策略在4.2.3介绍。

1. 禁忌表和禁忌准则

对于连续函数优化问题而言，仅仅禁忌4.1.1中所述的单个个体显然是没有意义的。由于函数为连续函数，这时需要根据函数特性重新定义禁忌准则。本文考虑使用禁忌目标函数值的邻域（此处邻域概念为 拓扑学概念，非上述TS算法邻域）来作为禁忌准则。

其具体实现为：设当前解的目标函数值为*f(x)*，若邻域点的目标函数值*f()*满足| *f()- f(x)*|≤*ε*，则*f()*被禁忌，其中*ε*为禁忌判断参数，具体数值在第五章实验部分描述。



图4.2 CTS-SS更新图例

图4.2介绍了CTS-SS的更新规则：红色菱形代表当前最优解，以该点为圆心，设定邻域半径后生成一个超球体邻域（图中实线圆框），在该邻域中存在无数个邻域候选解（灰色矩形），选取一定数目的邻域候选解与当前最优解进行比较（红色虚线）。若邻域最优解（红色矩形）优于当前最优解，且其不在禁忌表中（未被禁忌），则本次迭代获得了一个更优解，邻域最优解成为当前最优解，其余邻域候选解被舍弃。下次迭代将以本次得到的最优解为中心，再次生成新的邻域（虚线圆框），同时选择新的邻域（蓝色圆形）再次进行比较。

### 4.2.2 基于Sine混沌映射生成初始值

混沌理论是一种兼具质性思考与量化分析的方法，目前已经被广泛应用于各种动态系统（如：人口移动、化学反应等）中无法用单一的数据关系，而必须用整体和连续的数据关系才能加以解释和预测的行为。如果一个系统的演变过程对初始的状态十分敏感，那么可以称这个系统为混沌系统。一般而言，混沌系统具有非线性、遍历性、整体局部不稳定性、长期不可预测性等特点。目前混沌理论已被广泛应用于模拟随机数分布，用于代替随机取值。

目前比较典型的混沌映射有Logistic、Tent、Bernoulli Shift、Chebyshev及本文所用的Sine映射。需要说明的是，目前的工作[17]表明没有一种映射能在所有情况下都获得较好的寻优精度。在文献[6]中，作者已将Sine混沌映射应用于CNOP问题求解中，并取得了良好的效果。本文将结合GFDL CM模式初始的数据结构，将Sine映射应用于其初始值的生成中。

先前实验已表明，常规的随机初始化有很强的偶然性，而对单粒子算法而言，其能否快速收敛很大程度上依赖于初始解的选择。我们通过Sine映射，对个体进行初始化处理，其映射公式如下：

 （4.1）

 （4.2）

 （4.3）

 （4.4）

其中，*k*是当前迭代次数，*K*是最大迭代次数，*C*是个体（*Ci*代表第i个个体）。通过Sine映射随机生成个体*x0*，随后对其进行混沌计算，但迭代次数为*K*次时，迭代终止，这时*x0*便成为*xk*，同时将各个维度迭代后的值赋予*Ci*，这样就得到了一个初始个体。假设初始生成*pop*个初始个体，通过比较这*pop*个初始个体，从中选取拥有最佳目标函数值的个体作为算法的初始解。

 由2.2可以知道，CNOP中个体维度的取值值域为[-1,1]，而Sine函数的值域空间也为[-1,1]，所以我们无需通过任何过渡映射，便可以将Sine函数获得的初值应用于CNOP求解，其初值遍历性如图4.3所示。

图4.3 Sine映射生成初始值散点图

为了说明Sine混沌映射可以将原本聚集在一起的点重新分散，有助于避免出现局部最优。通过图4.4和图4.5在二维空间中验证了这一结论。



图4.4 二维空间中原本聚集在一个小区域内的点集（未经过Sine映射）

 图4.5 经过Sine映射后二维空间中的分散分布的点集

图4.4所示的是在横坐标与纵坐标区间均为[0,1]内一个点集，其中的每个点均集中在一个较小的区域内，这是为了模拟陷入算法局部最优的情况。图4.5则是图4.4中每个点经过Sine映射后各个点的分布情况。此次Sine映射的正弦函数迭代次数设为100，可以从图4.5中看出基本上每个点均已远离原来的位置，分布得非常分散，这说明通过混沌映射可以获得均匀分布的初始解，这样就可以最大限度地避免随机取值对结果带来的不确定性。

### 4.2.3 分段式搜索策略

由禁忌搜索的基本思想可知，CTS算法是一个逐渐逼近最优解的过程。随着迭代的进行，如果当前解已经接近最优解，这时通过集中搜索则可能更快收敛。然而，在传统的CTS中，邻域值和邻域生成的候选解个数都是恒定不变的，这为算法的改进提供了可能。

受到搜救过程的启发，我们可以将CTS的寻优过程分为两个阶段：第一阶段选取较大的邻域半径值和较少的邻域候选解个数，通过搜索的广度来逼近最优解，第一阶段结束后，进入第二阶段的搜索，第二阶段选取较小的邻域半径值和较多的邻域候选解个数，通过个体的集中来定位最优解，算法公式转化为4.5和4.6：

， *j*=1, 2, ... *n*; *i*=1,2,...,*m1*（4.5）

， *j*=1, 2, ... *n*; *i*=1,2,...,*m2*（4.6）

 其中第一阶段和第二阶段的邻域半径分别为*R1*和*R2*，邻域候选解个数分别为*m1*和 *m2*。整体的搜索阶段改进如4.6所示：

图4.6 CTS-SS搜索流程图

## 4.3 CTS-SS求解CNOP算法描述

基于上述CTS-SS的基本思想，可以得到其在CNOP求解中的算法描述：

|  |
| --- |
| 参数设置:  阶段1：候选解个数：*m1*；邻域半径：*R1*；最大迭代次数：*maxcycle1*；  阶段2：候选解个数：*m2*；邻域半径：*R2*；最大迭代次数：*maxcycle2*；  其他 ：适应度函数*f()*；特征空间维度*n;* 初始候选解个数：*pop;*  混沌迭代次数：*K*；禁忌判定参数：ε  *i=i1=i2=0,j=0; iter1=iter2=0; k=0*；禁忌判定参数：ε  映射混沌初始化：  1: for *i*=1 to *pop* do  2: for *j*=1 to *n* do  3: for *k*=1 to *K* do  4: 用公式（4.1-4.3）生成个体各个维度的值*Ci,j*;  5: end for  6: end for  7: end for  8: 使用公式(4.4)获取初始解*C.*  阶段1的运算：  9: for *iter1*=1 to *maxcycle1* do  10: for *i1*=1 to *m1* do  11: for *j*=1 to *n* do 12: 使用公式（4.5）计算*Si,j* ，设置*r* =*R1*  13: end for 14: 计算*f(Si)*  15: if *Si*不在禁忌表内then  16: if *f(Si)*-*best\_value*≧ε then  17: 将*C*列入禁忌表, *best\_value=f(Si)*  18: *C*=*Si*  19: end if  20: end if  21: end for  22: end for  阶段2的运算： 23: for *iter2*=1 to *maxcycle2* do  24: for *i2*=1 to *m2* do  25: for *j*=1 to *n* do 26: 使用公式（4.6）获取数值*Si,j*，设置*r* =*R2*  27: end for 28: 计算 *f(Si)*  29: if *Si*不在禁忌表内then  30: if *f(Si)*-*best\_value*≧ε then  31: 将*C*列入禁忌表, *best\_value=f(Si)*  32: *C*=*Si*  33: end if  34: end if  35: end for  36: end for  输出: 当前最优个体*C*及其适应值*best\_value* |

第五章将基于此设计数值实验，将其应用于求解GFDL CM模式CNOP，并研究其在ENSO最快增长初始误差中的应用。

# 5 实验与验证

## 5.1 实验环境

本文所采用的实验环境为国家超级计算中心广州中心“天河二号”超级计算机，所使用分区中CPU配置为Intel(R) Xeon(R) CPU E5-4640 0 @ 2.40GHz，使用RedHat 6.5 64位操作系统，GFDL CM2p1模式编程语言为Fortran，编译器为ifort11，实验程序编程语言为Java，jdk版本为1.8。

## 5.2 程序说明

本实验的完整过程分为两个模块：基于MATLAB的数据获取和降维模块以及基于Java的CTS-SS寻优算法模块。下面对这两个阶段的实现程序进行说明：

A．数据降维模块

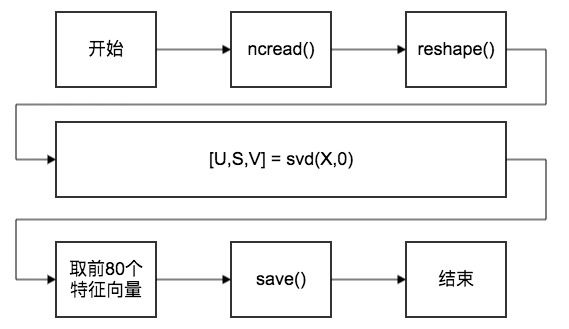
 在第一阶段，通过MATLAB自带的Netcdf工具对数据进行读取和选择，通过数据整合，获取和保存所有SST数据使用reshape函数获得将3.2中所述的样本数据集。之后通过svd函数对整体数据矩阵进行降维，取其前80个特征向量，使用save函数进行保存，即获得降维后的样本数据集。其主要运行流程和使用到的函数如下图所示，运行过程保存为一个.m可执行文件。

图5.1 数据降维阶段程序说明

B．CTS-SS模块

在第二阶段，Java程序首先通过Math库自带的Sine函数库生成一个80维的Sine特征向量，以此作为初始扰动矩阵。通过引入Java中的Netcdf库实现对.nc文件的操作，将第一阶段数据导入至内存中，完成数据读取工作。

然后通过引入Matrix库对数据进行矩阵乘法操作，获取初始扰动，这时结果是一个72000（360\*200）\*1的列向量，通过运用于第一阶段相同的reshape规则将该矩阵还原成一个360\*200的初始扰动矩阵，将该矩阵加入/INPUT文件中，即完成模式运行前准备工作。

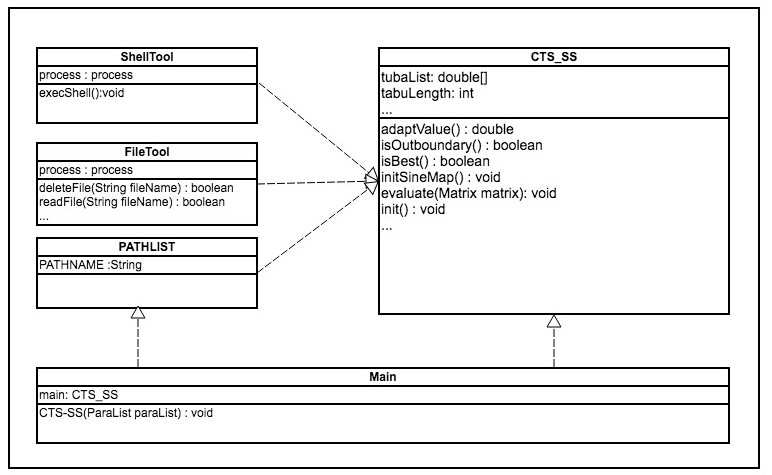
 之后通过运行Shell脚本执行模式，进入分阶段搜索。每次模式执行后都调用adaptValue函数获取当前结果，并通过isOutboundary来判断该结果是否还在数值范围之内。通过不断的迭代，寻找最优解并输出。该阶段整体程序结构如图5.2所示。

图5.2 CTS-SS寻优模块程序说明

需要说明的是，上述类图并不完整，忽略了一些具体实现的函数，只罗列了整体程序的运行框架。需要进一步说明的细节将在5.3描述。

## 5.3 实验步骤

本文基于如下流程图设计并进行数值实验：

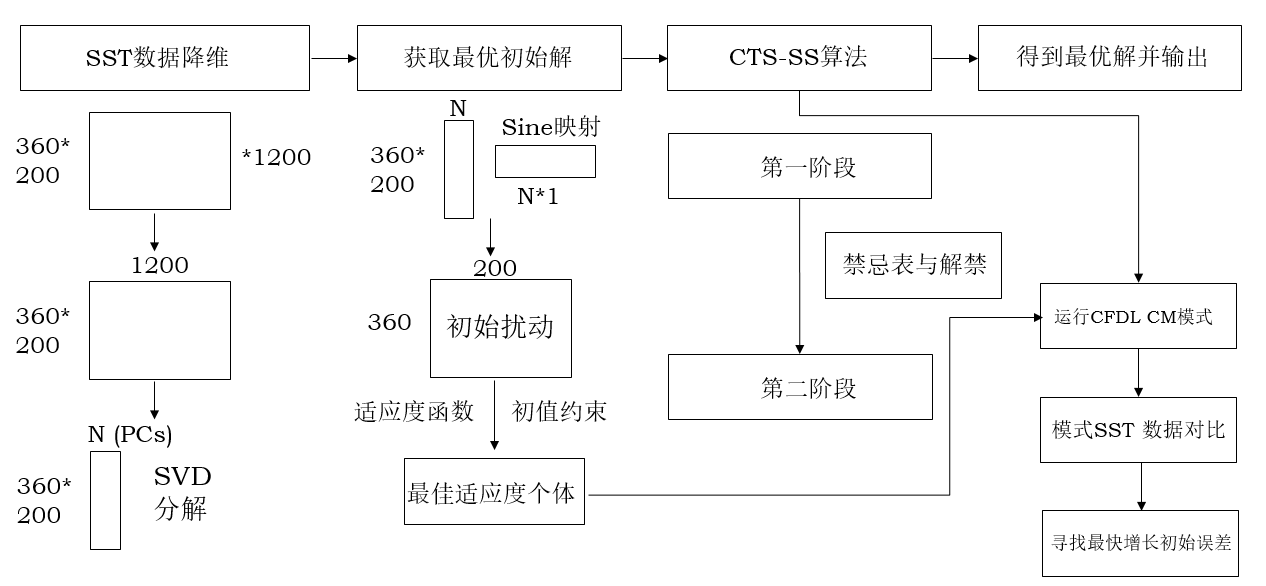


图5.3 数值实验整体流程图

下面分别描述具体实验步骤，值得一提的是，基于高内聚低耦合的程序设计原则，实验程序在完成后被打包为一个.jar文件，在使用时直接执行shell命令调用脚本即可。

### 5.2.1 获取SST数据及降维

实验过程基于完美模式假设，仅仅考虑SST对ENSO事件的影响。在第三章所述样本输入数据中，选择/INPUT文件夹下ocean\_temp\_salt.res.nc中SST作为主要输入衡量指标，将其作为训练集来生成初始样本。

通过运行控制实验100年（51年1月至150年12月），可以得到一个1200个360\*200的矩阵（划分粒度为月份，具体数值表示SST）。将这1200个矩阵生成用过映射关系组合成一个72000（360\*200）\*1200的初始矩阵，通过降维后，取其前80个特征向量作为主成分（具体原因参考3.3），并保存这部分数据。同时，保存100年模式运行结果的平均SST数据，将其作为模式基本场，整体的数据预处理步骤如图5.4所示。

图5.4 数据预处理流程图

### 5.2.2 获取最优初始解

在进行分阶段寻优过程前，需要配置程序主要运行参数，具体参数列表如下：

|  |
| --- |
| **private final int xAxis** = 360; *//水平分辨率* **private final int yAxis** = 200;  *//垂直分辨率* **private final int Dim** = 80; *//特征空间维度(主成分数目)* **private final int initNumber**; *//初始候选解数目*  **private final int tabuLength**; *//禁忌长度* **private int m1**; *//第一阶段候选解个数* **private int m2**; *//第二阶段候选解个数* **private int R1**; *//第一阶段邻域半径* **private int R2**; *//第二阶段邻域半径* **private int MAX\_CYCLE1**; *//第一阶段迭代次数* **private int MAX\_CYCLE2**; *//第二阶段迭代次数* **private double tabuParameter**; *//禁忌判定参数* **private double**[][] **tabuList** = **new double**[**Dim**][**tabuLength**]; *//禁忌表* **private double**[][] **initList**; *//候选初始解列表* **private double**[] **initValue** = **new double**[**Dim**]; *//初始解* **private double initEvaluation**; *//初始解扰动值* **private double**[] **bestValue**; *//当前最优解* **private double bestEvaluation**; *//最优解扰动值* **private double**[] **tempValue**; *//临时解* **private double tempEvaluation**; *//临时解扰动值* **private double**[] **localValue**; *//一次迭代中的局部解* **private double localEvaluation**; *//局部解扰动值* |

图5.5 实验主要配置参数

其中，模式分辨率和主特征数目（根据第3章实验结果，选取80个主特征）作为已知输入，而禁忌长度，初始候选解数目（由于GFDL CM模式规模和计算资源的限制，实际实验中选10个），迭代步长和迭代次数等参数可动态调节。

在进行寻优步骤之前，我们需要通过Sine映射生成多个候选初始解，并从中选取最优的一个作为模式初始解，其具体步骤如下：

1. 随机初始扰动生成

通过Sine映射生成一个80\*1的列向量，将其与5.2.1中获得的样本矩阵进行矩阵乘法，得到一个72000（360\*200）\*1的列向量，再根据相同的映射规则投影回原向量空间（360\*200），并将该初始扰动加入ocean\_temp\_salt.res.nc中SST。重复多次，得到多组初始随机扰动。

1. 寻找最优初始解

由于GFDL CM模式是一个动态运行的气候模式，只有在实验过程中获得及时输出的结果才有意义。故获得初始候选解后，立即运行GFDL CM2p1模式，由于模式采用Fortran语言编写，而模式运行完成时间并不确定，故采用Java语言编写一个轮询结构，每10分钟对shell脚本命令的输出进行字符匹配，以检验模式是否运行完成（同时将运行过程输出存储至日志文件中）。

|  |
| --- |
| *//调用shell* FileHelper.*exec*(**"bsub ./fr21.csh"**); *//轮询(10分钟一次),判断模式是否完成* **while** (**true**) {  **try** {  Thread.*sleep*(1000 \* 60 \* 10);  String temp = ShellHelper.*exec*(**"bjobs"**);  **if** (temp.equals(**""**)) {  **log**.write((**'\n'** + **"GFDL run finished!"** + **'\n'**).getBytes());*//一次新模式完成标志* **break**;  } **else** {  System.***out***.println(**"This cycle not finished! Waiting..."**);  }  } **catch** (Exception e) {  e.printStackTrace();  } } |

图5.6 轮询函数实现

检查到模式运行结束后，立即将该结果与5.2.1中所得到的模式态基本数据进行对比，通过适应度函数求出CNOP。

需要说明的是，对于得到的每一组初始扰动，其可能并不一定都满足扰动的初始条件，这时需要通过一个初值约束函数进行判断，需要判断其是否依然在超球体特征空间中，否则需要使用公式2.17将被投射到边界之外的点在重新投影回特征空间。通过比较这10个候选解的适应度函数值，从中选择最优的一个作为初始解，其运算流程如图5.7所示。



图5.7 获取最优初始解流程图

### 5.3.3 CTS-SS寻优与实验结果对比

在获得初始解后，进入分阶段搜索过程。通过预先设定的迭代次数和迭代步长，将寻优过程分为两个阶段。在每次迭代结束后，将当次迭代获取的最优解localEvaluation与之前迭代的最优解bestEvalutaion进行对比，若本轮迭代最优解优于当前最优解，则判断该解是否在禁忌表。若不在，进行交换，本轮最优解取代当前最优解，若在禁忌表中，则不替换。同时更新禁忌表，其主要实现如图5.8所示：

|  |
| --- |
| *//2.分阶段搜索* **for** (**curCycle** = 1; **curCycle** <= **MAX\_CYCLE1**; **curCycle**++) {  **for** (**int** i = 0; i < **m1**; i++) {  **if**(0 != times) { *//第一次不交换* **tempValue** = swap(**R1**, **tempValue**);  }  **tempSolution** = convert(**tempValue**);  evaluate(**tempSolution**, times);  times++;  }  **bestEvaluation** = isBest(**localEvaluation**, **bestEvaluation**); }  **for** (**curCycle** = 1; **curCycle** <= **MAX\_CYCLE2**; **curCycle**++) {  **for** (**int** i = 0; i < **m2**; i++) {  **tempValue** = swap(**R2**, **tempValue**);  **tempSolution** = convert(**tempValue**);  evaluate(**tempSolution**, times);  times++;  }  **bestEvaluation** = isBest(**localEvaluation**, **bestEvaluation**); } |

图5.8分阶段搜索实现

### 5.3.4 得到最优解并输出

在完成搜索过程之后，我们将得到的结果使用NCL软件进行图像绘制，比较其在太平洋区域SSTA的变化，与事前确定的作为参考态的El Niño事件进行对比（具体在5.4实验结果分析中说明）。若结果为本次寻优过程的最大CNOP值，且通过理论和图像分析，认为该结果有较为明显的数值特征和物理发展意义，则可以认为本次结果为满足条件的最快增长初始误差，将其作为结果输出，实验结束。

## 5.4 实验结果分析

为了寻找最快增长初始误差，需要先找基本场的ENSO事件，100年控制实验，得到模式运行结果与实际观测值的对比（图5.9），可以看出68/69年,79/80年发展为较为典型的ENSO事件，将其与97/98年的观测值进行对比可以看出，68/69年具有明显向西太平洋发展的趋势，79/80年与97/98年观测值拟合较好，实验中选取68/69年作为模式初始场。



图5.9 ENSO事件SSTA对比（上图为控制实验68/69年，中图为控制实验79/80年，下图为98/99年观测值）

经过初始扰动的叠加和CTS-SS寻优，算法寻找到一个较为典型的最快增长初始误差。下面借助NCL画图工具，将未叠加扰动的结果和叠加最优初始扰动的结果分别与模式态100年SST相减得到一年海表面温度异常（Sea surface temperature anomaly，简称SSTA）数据及发展，并根据季度进行分析。

|  |  |
| --- | --- |
| C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\39V{0]GA83M1)WA)94ITW}Q.PNG | C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\1.png |
| C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\o2.png | C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\2.png |
| C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\o3.png | C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\3.png |
| C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\s.png | |

图5.10 第一季度数据SSTA对比（左侧为未加扰动，右侧为叠加扰动）

对比第一季度SSTA的结论如下：

（1）在积分的前三个月中，中高纬度地区海表面温度异常基本无明显变化，扰动并未出现明显的物理特征，但在叠加扰动后模式积分的第三个月中，东太平洋的正向扰动明显增加；

（2）随着扰动的发展，南半球中低纬度太平洋正向SSTA出现减弱趋势。

|  |  |
| --- | --- |
| C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\o4.png | C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\4.png |
| C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\o5.png | C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\5.png |
| C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\o6.png | C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\6.png |
| C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\s.png | |

图5.11 第二季度数据SSTA对比（左侧为未加扰动，右侧为叠加扰动）

对比第二季度SSTA的结论如下：

（1）在第二季度中，热带太平洋未叠加初始扰动时，表现出东太平洋负向异常，西太平洋出现较为明显的正向异常；而叠加初始扰动后，在东太平洋表现出较大范围的正向异常，西太平洋则表现出负向异常；

（2）在南半球中低纬度太平洋地区与第一季度表现相似；

（3）东太平洋的正向异常，逐渐向西北方向传播，至北回归线附近，随即沿西南方向向西太平洋传播。

|  |  |
| --- | --- |
| C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\o7.png | C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\7.png |
| C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\o8.png | C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\8.png |
| C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\o9.png | C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\9.png |
|  | |



图5.12 第三季度数据SSTA对比（左侧为未加扰动，右侧为叠加扰动）

对比第三季度SSTA的结论如下：

（1）在赤道太平洋区域出现了明显的正向异常，尤其是在八九月份在东太平洋，未叠加扰动时，模式结果会出现较为明显的负向异常，而叠加扰动后，模式结果在赤道太平洋区域出现明显的全场的正向异常；

（2）与第二季度类似，叠加扰动后出现的正向异常传播至西太平洋后向东发展，并在中东太平洋区域发展；

（3）在未叠加扰动时，东太平洋开始出现的正向异常在向西太平洋传播的过程中能量不断耗散，最终在西太平洋消失。

|  |  |
| --- | --- |
| C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\o10.png | C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\10.png |
| C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\o11.png | C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\11.png |
| C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\o12.png | C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\12.png |
| C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\s.png | |

图5.13 第四季度数据SSTA对比（左侧为未加扰动，右侧为叠加扰动）

对比第四季度SSTA的结论如下：

（1）叠加扰动后，中纬度东太平洋边界正向异常不断向西太平洋传播，而未叠加扰动赤道东太平洋区域始终表现出负向异常；

（2）综上可以推测，气候信号由东太平洋产生，沿边界向北传播，到达北回归线附近，再沿西南方向向西太平洋传播，到达西太平洋后，叠加扰动后，该信号会向东继续传播，而未叠加扰动时，信号能量在传播过程中耗散，在西太平洋消失，没有产生向东传播的信号。

本文同时将全年平均态SSTA进行对比：

|  |  |
| --- | --- |
| C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\oss.png | C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\ss.png |
| C:\Users\Kris\AppData\Local\Microsoft\Windows\INetCache\Content.Word\s.png | |

图5.14 全年平均SSTA对比（左侧为未加扰动，右侧为叠加扰动）

全年平均SSTA分析如下：

（1）叠加扰动前，赤道太平洋区域普遍表现为负向异常，叠加扰动后，在中东太平洋区域出现正向异常。

（2）叠加该扰动，在模式积分过程中出现了与厄尔尼诺事件近似的发展过程，基本认为这是本次实验中的发展最快的初始误差。

需要说明的是，由于实验运算资源的限制和GFDL CM模式运行时间过长两方面的原因，本实验过程只从有限的初始误差中寻找到了一个潜在发展为最快增长初始误差的扰动，但从改扰动的生成和发展可以看出，CTS-SS算法在寻找目标函数时体现了良好的性质，将其运用于GFDL CM模式CNOP问题的求解是可行的。

# 6 结论和展望

## 6.1论文总结

本文将CTS-SS算法应用于GFDL CM模式中CNOP求解，并设计了数值实验初步将其应用于寻找ENSO事件中最快增长初始误差。对于发展成ENSO事件的初始误差，通过图像比较、数值处理等手段进行衡量，最终确定发展成为最快增长初始误差的最优扰动。实验结果验证了此方法的可行性，为GFDL CM模式提供了更为广阔的应用前景。

## 6.2 进一步研究方向

在现有实验的基础上，之后还有许多需要完善的工作，主要有如下几个方面：

后续研究：在两类影响ENSO事件的发展过程中，最快增长初始误差影响ENSO的预报过程，最优前期征兆影响其发生过程，通过这两个研究内容确定目标观测敏感区，设计并进行敏感性实验，进一步验证最快增长初始误差，找到目标观测敏感区。其中，使用智能算法寻找最优前期征兆的工作由本课题组秦博完成。后续工作可将这两种现象得到的SSTA进行对比，获得ENSO事件的目标观测敏感区，这对ENSO事件的研究将有更重要的意义。

并行处理：GFDL CM作为全球海气耦合模式，每一次运行积分的时间开销都是不可忽略的（这一点不同于某些模式，如MM5等，可以在较短时间内得到结果）。而智能算法（如本文采用的CTS-SS）的寻优过程往往是成百上千步，这样每次实验的时间非常巨大，通过增加运行时GPU数目，引入GPU并行计算等手段，可以提高模式运行效率。

参数调整：通过算法参数的调整，提高寻优效率。这一步还需要大量的实验工作，寻找到最优的实验参数，以寻找到求解GFDL CM模式的最佳配置参数。

扩展思路：智能算法作为人工智能的一种体现，其主要思想是逐步获取最优解的过程。基于相似的基本原理，机器学习，深度学习等理论中的思想从理论上来说也可以应用于海气耦合模式中最优化问题的求解过程中，这是另一个可能的求解方向。

# 参考文献

[1] Mu M., Duan W. S. & Wang B., Conditional nonlinear optimal perturbation and its applications. Nonlinear Processes in Geophysics, 2003, 10: 493-501

[2] 穆穆，段晚锁.ENSO 可预报性研究的一个新方法:条件非线性最优扰动.科学通报, 2003,48(7):747-749

[3] 陈磊. ENSO最优前期征兆和最快增长初始误差的相似性及其在识别ENSO目标观测敏感区中的应用[博士学位论文].北京：中国科学院大学, 2015

[4] 徐辉. Zebiak-Cane ENSO 预报模式的可预报性问题研究[博士学位论文].北京:中国科学院大气物理研究所,2006

[5] 温仕成. 求解CNOP的算法优化及其应用[博士学位论文].上海：同济大学, 2015

[6] Yuan S., Qian Y., Mu B. Paralleled Continuous Tabu Search Algorithm with Sine Maps and Staged Strategy for Solving CNOP. In: Wang G., Zomaya A., Martinez Perez G., Li K. (eds) Algorithms and Architectures for Parallel Processing. Lecture Notes in Computer Science, vol 9530. Springer, Cham，2015

[7] 孟佳佳，杨宇星. GFDL模式对太平洋海表面温度的年际和年代际变率的模拟评估.海洋科学, 2016

[8] 胡一波. 求解约束化问题的几种智能算法[博士学位论文].陕西：西安电子科技大学, 2009

[9] Mu B, Zhang L. L.: PCAGA: Principal Component Analysis Based Genetic Algorithm for Solving

Conditional Nonlinear Optimal Perturbation. Submitted to International Joint Conference on Neural

Networks , 2015

[10] Zhang L L, Yuan S J, Mu B, et al. CNOP-based sensitive areas identification for tropical cyclone adaptive observations with PCAGA method. Asia-Pacific J Atmos Sci, in press, 2016

[11] Chen L, Duan W S, Xu H. A SVD-based ensemble projection algorithm for calculating con-

ditional nonlinear optimal perturbation . Science China: earth sciences, on press

[12] 穆穆，王强. 条件非线性最优扰动法在大气与海洋目标观测研究中的应用.气象学报，2014

[13] Jonathon Shlens. A Tutorial on Principal Component Analysis. Google Research，2014

[14] F.Glover. Tabu search-Part 1, ORSA Journal on Computing,vol.1.no.3, 1989

[15] F.Glover. Tabu search-Part 2, ORSA Journal on Computing,vol.2.no.1, 1990

[16] 王明兴. 连续禁忌搜索算法改进及应用研究[硕士学位论文].浙江：浙江大学，2005

[17] 钱一闻. 并行化智能算法求解CNOP研究及其在ZC和MM5模式中的应用[硕士学位论文].上海：同济大学，2015

[18] Xin Z. Research on optimization performance comparison of different one dimensional chaotic maps[J]. Application Research of Computers, 2012, 3: 032.

[19] 穆穆，王强.非线性最优化方法在大气-海洋科学研究中的若干应用.中国科学，2017

# 谢辞

时间乃最大的革新家，而人生就是用时间换取自我革新与成长的过程。我们能留下的，终归只有已经尽心尽责的满足感，以及拼尽全力的证据。

本科四年时光飞逝而过，我将要踏上一段新的革新之路，在完成毕业设计这颇有几分仪式感的时刻，我想要对很多人表达我最诚挚的谢意。

首先要感谢我毕业设计的指导老师袁时金老师，袁老师认真负责的工作态度，对学生孜孜不倦地教导，都使我受益良多。没有袁老师的指引， 我不可能完成本篇论文的写作。

我要感谢穆斌老师，穆老师严谨的治学态度，从容的处世哲学和永不满足的学术追求，都深深地影响了我。办公室那盏每晚12点前几乎不会熄灭的亮灯更是指引我不断追求上进的动力。所谓大师在前，不敢止步。

感谢软件学院的许多老师，和你们的邮件交流或面谈都让我如沐春风，你们让一个困惑的年轻人从无所事事、得过且过到开始真正思考如何把一些事情做好，感谢你们。

感谢那些一起熬夜写项目，一起通宵复习的小伙伴们。那种在一起完成各种课程设计，被一个个Bug弄得焦头烂额，不得不加班加点的情形还历历在目。事后想起来，这些居然成大学最美好的回忆。那种纯粹对编程的热爱和仅仅想把事情做得更好的些许“匠人情怀”，就是我们留在这行继续摸爬滚打的最好理由吧。

最后，我要深深感谢我的父母，是你们让我有机会去践行自己的目标，实现自己的价值。